

DOCUMENT RESUME

ED 190 643

TM 800 438

AUTHOR Dyer, Henry S.
TITLE Parents Can Understand Testing.
INSTITUTION National Committee for Citizens in Education, Columbia, Md.
SPONS AGENCY Babcock (Mary Reynolds) Foundation, Inc., Winston-Salem, N.C.
REPORT NO ISBN-0-934460-08-6
PUB DATE 80
NOTE 103p.
AVAILABLE FROM National Committee for Citizens in Education, Suite 410, Wilde Lake Village Green, Columbia, MD 21044 (\$2.50)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Accountability: Elementary Secondary Education: *Information Needs: *Measurement: Minimum Competency Testing: *Parent Participation: Scores: Test Bias: *Testing: Testing Problems: *Test Interpretation: Test Validity

ABSTRACT

This handbook aims to help parents become more familiar with the testing that goes on in their community schools. The first two chapters provide general background on testing, giving the reason and scope of educational testing today, and highlighting the fundamentals of the testing-measurement process. Chapters 3 and 4 consider the various pitfalls to be avoided: common misconceptions about tests and scores; and matters prominent in current discussions, such as test bias and teacher accountability. In the last chapter, a series of questions are raised that parents typically ask of teachers and administrators. Included are four appendices: Appendix A is a tabular summarization of state activity in minimum competency testing during 1972 through 1978; Appendix B is a glossary of terms (these terms are followed by an asterisk in the text itself) specific to the educational testing field; Appendix C lists selected readings concerning testing, and Appendix D reproduces Public Interest Principles for the Design and Use of Admissions Testing Programs. Also included is a list of available publications of the National Committee for Citizens in Education. (Author/GK)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED190643

- TM 800438

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

parents can understand testing

"PERMISSION TO REPRODUCE THIS
MATERIAL IN MICROFICHE ONLY
HAS BEEN GRANTED BY

J. Rioux

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

National Committee For Citizens in Education

The National Committee for Citizens in Education is a nonprofit, tax-exempt membership organization dedicated to increasing citizen involvement in the affairs of the nation's public schools. NCCE is a successor to the National Committee for Support of the Public Schools, founded in 1962. In 1973 the Committee was reorganized, took its new name and reconstituted its purpose. Since then NCCE:

- Published a major report, *Children, Parents and School Records*, which received national attention in *Parade Magazine*, *Time Magazine*, *The New York Times* and on ABC Television.
- Supplied vital information incorporated in the Family Educational Rights and Privacy Act of 1974, which became law November, 1974.
- Conducted hearings across the United States on who controls America's public schools, taking testimony from hundreds of individuals and organizations.
- Joined with other national organizations to convene a National Conference of Title I Parents.
- Convened the first national meeting of active local parent groups held since 1910.
- Established the first public-interest, toll-free telephone hotline, 800-NET-WORK, to help parents get rapid information and help.
- Established *The Parents' Network*,* to mobilize citizens for action to improve our nation's public schools, and to support existing local parent/citizen groups.
- Established the Citizens' Training Institute which trains parents and citizens in ways that will enable them to work more effectively in schools.
- Established NETWORK, a newspaper for parents about schools.
- Produced a series of handbooks on topics rated as priority concerns by parents.
- Established a citizen oversight program. Local level monitoring of federal laws meant to be of assistance to public school children and their parents. Family Educational Rights and Privacy Act—Education of All Handicapped Children Act.
- Initiated special projects designed to explore problems and create new access points for parents in influencing decisions affecting the quality of public education which children receive. School Based Management—Single Parents and the Public Schools.

* Over 325 local groups—combined membership of 200,000 persons.

© 1980 National Committee for Citizens in Education

Library of Congress Catalog No. 80-80939

ISBN-0-934460-08-6

Cover Design—Patricia Silbert

About the Author

Henry S. Dyer first became interested in testing more than forty years ago when, as an elementary school teacher, he persuaded his fellow teachers that it might be useful to give their pupils some standardized achievement tests. Ever since then he has been trying to get teachers and others, including parents, to understand what testing is all about and how it can and cannot contribute to the work of the schools. This effort eventually took him to Harvard, where, among other things, he studied educational measurement and later taught it there to prospective teachers and to some of his colleagues on the Harvard faculty.

Since then he served for a time as an Associate Director of the College Entrance Examination Board and more recently as a Vice President of Educational Testing Service. In both spots he continued to work at spreading the truths about testing so that the use of tests in the schools and elsewhere might become more cautious and more educationally constructive. Now in retirement, he is still trying to figure out how best to get the same message across.

Acknowledgments

This handbook has benefitted from the attention and careful review of the following people: Chrissie Bamber, Deborah Clayman, Janice Gygi, Jim Harwood, Mitchell Lazarus, Mike Matthias, Carl Marburger, Beverly Norbeck, Shirley Oehling, Bill Rioux and Stanley Salett.

In addition to these readers of the author's final manuscript, Mr. Dyer is indebted to William H. Angoff, Eleanor V. Horne and Jerome A. Weisbrodt for reading and commenting in considerable detail on a draft of the entire work and to Lawrence A. Babb, Gerry Ann Bogatz and Richard M. Jaeger for reading and commenting on portions of that manuscript in its early stages.

The author and staff of NCCE wish to give special credit and thanks to editor Marian Hoffman for helping to present a mass of technical material in what we believe to be a highly readable and useful form.

We are grateful to the Babcock
Foundation for its support in
the development of this publication

Contents

	<i>Page</i>
Introduction	v
Chapter 1	
Highlights of the Testing Scene	1
Why Schools Give Tests	1
The Kinds of Tests Schools Give	5
The Tester Network	7
Some Controversial Issues	9
Chapter 2	
Educational Testing as Measurement	13
Elements of the Testing Process	13
Translating Test	
Performance Into Numbers	15
Reliability and Validity	18
The Reliability Question	18
The Validity Question	20
Chapter 3	
Pitfalls in Communication	25
Norms	25
Grade Equivalent Scores	27
Mental Age	29
IQ and Aptitude	29
Over- and Under-Achievement	31
Minimum Competence	32
Chapter 4	
Policy Matters	35
Bias in Testing	35
Testing for Educational	
Accountability	38
Testing for Minimum Competence	40
Chapter 5	
Questions to Ask	
About Local Testing Matters	47
Questions About the	
Teachers' Own Tests	48
Questions About Tests Purchased	
From Test Publishers	50
Questions About Tests Supplied	
by the State Education Authority ...	53
Questions About Miscellaneous	
Local Testing Matters	55

NCCE/iii

<i>Appendix A</i>	
State Activity in Minimum Competency Testing	57
<i>Appendix B</i>	
Glossary of Terms	71
<i>Appendix C</i>	
Selected Readings	81
<i>Appendix D</i>	
Public Interest Principles for the Design and Use of Admissions Testing Programs	84

Introduction

From the first grade through high school, a child puts in a great many hours taking tests. The educational impact of all this testing can vary enormously from child to child and from school to school. Whether the testing is in fact helpful or harmful, useful or useless, depends on the nature and quality of the test itself, the manner in which it is administered, scored, and interpreted, and the uses to which the results of the test are put.

Because of this, parents of school children should become more familiar with the testing that goes on in their own communities. How does it and should it contribute to their children's education? What effects does it and should it have on how teachers teach and on how the schools are run? The aim of this handbook is to help parents find intelligible answers to such questions.

The first two chapters provide some general background on testing. Chapter 1 gives an idea of the scope of educational testing today, takes a look at the kinds of people involved in it, and touches on the controversies created by it. Chapter 2 deals with some fundamentals of the testing-measuring process that are sometimes forgotten by the testers themselves and are often overlooked by the test-using public.

The next two chapters consider various pitfalls to be avoided. In Chapter 3 some of the common misconceptions about tests and test scores are discussed, and in Chapter 4 a number of matters that figure prominently in current discussions of testing—such as the problem of test bias and the role of test results in holding teachers and schools accountable for student progress—are examined.

The last chapter presents a series of questions that parents might ask of teachers and administrators in the schools. The hope is that these questions will provide an agenda for a fruitful discussion of local testing operations, with a view to maximizing their utility as part of the overall educational process.

Educational testing is an extremely complex

subject, and some of the language associated with it is admittedly difficult to understand. We've used a number of technical terms throughout this book because you'll need to be familiar with them when you discuss local testing matters with school officials. An asterisk appearing after a word or phrase is used the first time means that you'll find a definition of it in Appendix B, the Glossary of Terms.

Highlights of the Testing Scene

Why Schools Give Tests

Schools test students for a variety of reasons, some of which are quite clear to the testers and the children, some of which are known only to the testers. Some of the reasons may even be obscure to all the parties involved, including students, teachers, administrators, and parents.

How clearly the purpose of any test is understood depends a good deal on where the test comes from and on who has called for it to be given. When a classroom teacher prepares a test for his or her students, both the teacher and the students are sure to know why the test is being given. When a state education agency orders all schools to give the XYZ Reading Test to all pupils, there is bound to be some confusion about what the testing is for.

Why then do schools test students? Here are ten of the most common reasons:

1. To help motivate student learning: A teacher may say something like this to a class: "Next week I'll be giving you a test on everything we've covered this past month. It'll be a tough test, so I advise you to review the material thoroughly." Or this: "Here are some questions like those you will be up against on the College Boards, so you'd better get this stuff down cold!" There are, of course, many subtler ways to use tests as incentives to study. Some people feel that this use of tests encourages learning for the wrong reasons. But the belief is widespread that it works and is, therefore, educationally indispensable.

2. To individualize instruction: For the past twenty years, there has been considerable growth in the use of what are called mastery tests* or criterion-referenced tests.* For the most part, these tests have focused on the learning needs of ele-

mentary school pupils. The results are supposed to tell the teacher what specific reading and math skills each child has already mastered, what skills he or she still needs to master, and what particular exercises he or she should work on next to make progress. In this system, the tests tend to be short, individually administered, and frequent, so that each child can move ahead as rapidly as possible. A less systematized approach to individualized testing occurs when a teacher gives a class an essay test,* grades each student's paper, and covers it with comments suggesting in detail what the student should do to improve the writing. Whatever the method, the use of tests for individualizing instruction has a strong appeal. It is, however, so demanding of teacher time—especially in large classes—that it is not feasible unless the teacher has the help of trained aides.

3. To help make placement decisions about students: This is probably the most common reason for giving tests *en masse*, and it has a lot of history behind it. In 1865, for instance, the state education authority of New York established the Regents Examinations to determine which students should be admitted to high school. In 1901, the College Boards were created to help colleges select students for their freshman classes. Around 1920, schools began using IQ tests* to sort pupils into classes for the "gifted," the "normal," and the "educable mentally retarded." In the 1930s, reading readiness tests were introduced to see which students should be admitted to first grade. Today, minimum competency tests* are being required by a growing number of states to determine who shall be promoted from one grade to the next and/or who shall receive a high school diploma. From the start, the use of such tests to make summary decisions about students has provoked intense debate.

4. To help students make decisions about themselves: Some years ago, the American School Counselors Association adopted the policy that the guidance counselor should "assist the pupil to meet the need to develop personal decision-making competency" and should "assist each pupil . . . to understand his aptitudes, interests, attitudes, and opportunities for self-fulfillment." In other words, school guidance counselors are supposed to use test results in ways that will help students figure out their own educational and career

goals. This ideal is generally applauded. However, it is difficult to attain, since guidance counselors with the necessary expertise are in short supply, and since so many test scores are all too easily misinterpreted by counselors, students, and parents alike.

5. To compare the effectiveness of schools:

Not so long ago the usual method of measuring a school's educational effectiveness consisted of examining the classroom facilities, teachers' credentials, pupil-teacher ratio, up-to-dateness of textbooks, and the like. More recently, the assessment of a school has depended on the scores its students make on tests—especially standardized achievement tests.* Thus, if the average reading score of sixth graders in School A is well above the average score of sixth graders in the nation, it is inferred that School A is doing an effective job teaching reading. And if the sixth-grade reading scores in School B average much less than those of sixth graders nationwide, it is inferred that School B's reading instruction is inept. This approach to the measurement of school effectiveness has been vigorously attacked for being technically unsound and so simplistic as to give a false impression of how well or badly any school may actually be serving its students.

6. To evaluate the results of experimental programs: Whenever a new method of teaching or a new course is introduced into a school, the question arises whether the innovation aids student learning. The answer to this question is often sought by comparing the test results of students in the experimental program with those of students in the regular program. Often the results of the testing are inconclusive and debatable, but the testing still goes on in the hope that someday there will be clear answers to the question: What works best for which students under what educational conditions?

7. To provide for continuity in the educational experience: As students go from grade to grade and from school to school, they come under the care of a succession of teachers, each of whom needs information about the past accomplishments of the incoming students in order to plan instruction for the upcoming year. To meet this requirement, most schools now maintain a cumulative record* on each student, which shows, among other things, the results of tests the stu-

7

dent has taken in past years. The test record is especially important for the proper placement of the many thousands of students who transfer each year from one school system to another.

8. To cooperate with test publishers in the try-out and norming of new tests: When developing a test for general use in the schools, a test publisher ordinarily seeks two types of information: (1) information to point up technical weaknesses that need to be corrected before the test is put in final form; and (2) information needed for establishing national norms* (the average scores for a defined population of students; for example, all sixteen-year-olds or all seventh graders) *after* the test has been put in final form. To meet the first need, the publisher gets schools to give early drafts of the tests to their students, so that imperfect items can be located and revised or eliminated, and so that teachers can report back on the appropriateness of the tests for their students. To meet the second need, the publisher must have the test administered to a representative sample of schools and students all over the country.

9. To cooperate with national testing agencies in the administration of tests to be used by various institutions for the selection of students: In this case the school, usually a high school, provides the facilities and the trained personnel for giving tests to students in the community who are seeking admission to an institution that requires those tests. The principal benefit to the student is that it makes the testing available in familiar surroundings close to home. It also gives the school's staff a chance to gain useful experience by working directly with personnel from national testing programs.

10. To comply with the official requirements of state and federal agencies: Such requirements came into national prominence with the passage by Congress of the Elementary and Secondary Education Act of 1965. This act included the following clause: "That effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally deprived children." The execution of this order was left to the states and the local school districts, many of which have had considerable difficulty meeting the requirement.

At least one effect of the act, however, has been the spread of a variety of statewide testing and assessment programs designed to provide statistics on pupil achievement district by district and school by school. As a consequence, there is no doubt that some schools give tests for no deeper reason than that they are ordered to do so by higher authority. In such circumstances, the quality of the statistics and the usefulness of the test results at the local school level may be questionable, since students and teachers may have some trouble understanding their own roles in the test operation.

As you can see, educational testing has many purposes, some of which may not have much to do with the education of the children who take the tests. Clearly, educational testing will have a better chance of contributing to your child's education if more people, both inside and outside the schools, understand the reasons for giving any test. This can be done by persistently asking school personnel in your own district two tough questions—one before and one after the testing takes place: (1) Why and in what ways will the giving of this test help your child, either directly or indirectly? (2) How specifically has the giving of this test actually contributed to your child's education, either directly or indirectly?

Getting satisfactory answers to these questions should be your first order of business if you are concerned about the quality of your child's education. For until the answers are forthcoming, much of the testing that goes on in schools is likely to be less useful than it could or should be.

The Kinds of Tests Schools Give

The tests schools give carry so many different labels and have so many different shapes and uses that it is easy to become confused by the whole operation. To get your bearings, it is helpful to divide all educational tests into general categories defined (1) by what student characteristics (known as attributes) the tests are supposed to measure, and (2) by the way the results are most often used.

The two main kinds of student attributes measured are: cognitive attributes* and affective attributes.* The cognitive attributes have to do with the students' knowledge; the affective attributes

have to do with their feelings, their likes and dislikes, and their beliefs.

Similarly, the results of a test are categorized according to whether they are primarily intended to measure what the students know or feel at the present time, or to predict the students' future status.

Thus we have four broad types of educational tests:

Type I. Tests that measure a child's current knowledge (cognitive attributes) in various areas of study. These include local teacher-made tests, tests of mastery, criterion-referenced tests, standardized achievement tests, statewide assessment tests, and tests of minimum competence.

Type II. Tests that are used to predict how well a child will do in the future in various areas of study. These include general intelligence tests, scholastic aptitude tests, tests of specific aptitudes (clerical, mechanical, musical, etc.).

Type III. Tests that measure a child's feelings, likes and dislikes, and beliefs. These include tests of self-esteem, tests of ethical and social values, attitude scales,* and personality tests.*

Type IV. Tests that are used to predict what a child will be interested in and care about in the future. These tests include academic and vocational interest inventories.*

The boundary lines between the test types are not hard and fast. An achievement test in reading, for instance, may serve just as well as a verbal aptitude test to predict how a student will do later on in school subjects that involve a lot of reading. Also, it is sometimes hard to decide whether certain test exercises are supposed to be measuring cognitive or affective attributes. Consider the following two exercises, one of which comes from a personality test and the other from a standardized achievement test in writing:

1. As you look back, you can undoubtedly find one or more points of decision where you took one path rather than another. The number and significance of these decisions will vary with your age, experience, and degree of parental control. Describe such a decision and indicate the influences—person, book, experience, event—which prompted you to make the choice you did.

2. Here is a picture. Your task will be to make up as dramatic a story as you can about it. Tell what has led up to the event shown in the picture, describe what is happening at the moment, what the characters are feeling and thinking, and then give the outcome.

In this case the first exercise happens to come from an achievement test in writing and the second from a personality test.

Even though there are not clear boundaries setting off the four test types, the classification scheme can be useful in getting a rough approximation of the kinds of testing that are emphasized in your local school. There is little question that, in the nation at large Type I testing predominates, since most teacher made tests are of that type. But there are exceptions—schools whose teachers are deeply committed to the humanistic approach to education may give more Type III tests than any other kind.

The Tester Network

In the United States during the past seventy-five years or so, there has been considerable growth in the number of organizations concerned with the advancement of educational testing as a science, an art, and a commercial enterprise. A recent count shows that there are now no fewer than 130 publishers that supply tests for use in American schools. Most of these are small operations specializing in the production and distribution of one or a few little-known tests. But there are nine that have dominated the field; a few of them are non-profit operations like Educational Testing Service and the Education Commission of the States, but most of them are profit-making concerns like California Test Bureau¹ and Science Research Associates.²

Working with the test publishers are the actual test makers. These are the men and women who plan the tests, create the exercises that go into them, and check them out in the schools to see if they are working as they should. Some of these test makers are on the publishers' staffs, but for the most part they are outside consultants—often

¹A subsidiary of McGraw-Hill.

²A subsidiary of International Business Machines.

professors of education or psychology—who work under contract to develop tests related to their specialties.

Most of the tests supplied by the publishers are sold directly to the school districts to be used when and how the local authorities decide. Usually the local testing operation is supervised by someone with a title like Director of Evaluation and Guidance, or Assistant Superintendent for Research. His or her job is to see that the tests are properly administered in each school building and to ensure that the test results are properly interpreted by teachers, students, school principals, and parents.

In addition, the local supervisor may be designated to handle the administration of a number of so-called "secure testing programs." These are the ones where all test booklets are kept under wraps until students open them in the examination room and then, when the test session is over, are promptly returned to the test supplier before teachers or other school personnel can get a peek at them. College entrance tests are in this category, as are some of the tests required by state education authorities.

Test specialists are also found in various federal and state agencies, such as the U.S. Department of Education and the research and evaluation divisions of most state departments of education. Some of them work at the headquarters of such organizations as the American Council on Education, the National Education Association, and the American Federation of Teachers.

Over the years, the members of these various groups of testers have formed a number of professional organizations to facilitate communication among themselves. Probably the two organizations that try hardest to bring together the test makers who work for the suppliers and the test givers employed by school districts are the National Council on Measurement in Education (NCME) and the Association for Measurement and Education in Guidance (AMEG). Other professional groups include the Psychometric Society (PS), certain divisions of the American Psychological Association (APA), the American Personnel and Guidance Association (APGA), and the American Educational Research Association (AERA).

Although the interests, as well as the memberships, in these organizations overlap to a degree,

some of them appeal primarily to individuals who deal almost exclusively with test theory, while others appeal primarily to those concerned with the nuts and bolts of testing in the schools.

Despite this, there is at least one continuing effort to combine forces in order to establish basic principles to which the production and use of tests should conform. Since the 1950s, representatives of APA, AERA, and NCME have been getting together periodically and producing a document listing standards for the production and use of tests. Their latest document, published by APA in 1974, is called *Standards for Educational and Psychological Tests*. A revision is now in the works.

Setting standards provides a kind of conscience for those test makers and test users who are able to understand them, but it does not guarantee that the standards will be understood and adhered to by a lot of other people, who simply take the numbers that the tests generate at face value. Among these is the group that forms the crucial center of the testers' network—the classroom teachers, who week in and week out are giving millions of tests in schools all over the country. Not much is known about how close the teacher-made tests come to meeting any kind of quality standards, but the evidence that exists is somewhat discouraging. Two surveys in 1967 found that the great majority of teachers had never had more than a smattering of instruction in how to prepare tests for classroom use, or in how to evaluate the tests that publishers supply and how to make appropriate use of them. Some professors of educational measurement have tried to correct this situation, but as far as anyone can tell they have not yet made much headway.

Some Controversial Issues

One reason that professional testers have trouble getting their messages across to classroom teachers and to the public at large is that they tend to exhaust themselves in deep debates with one another. A second reason is that the language they use is often incomprehensible to the majority of test users.

There are six issues that seem to be the most controversial, not only among the testers themselves but also among school people and their constituencies. In broad outline, the pros and cons go as follows:

1. Whether or not tests are biased: For at least fifty years, charges have been made that the tests schools give are unfair to children from so-called disadvantaged backgrounds. IQ tests in particular are said to mislabel children whose native language is not standard English as being mentally deficient. The defenders have claimed that, on the contrary, the tests have democratized education by spotting a lot of gifted children from all walks of life and giving them educational opportunities they would otherwise have missed. The argument has been confounded by the related question whether intelligence, as measured by tests, is primarily due to the child's biological inheritance or to the environment in which he or she has grown up.

2. Whether or not tests freeze the curriculum: It is sometimes charged that when certain tests are given to all the students in a state or a school district, it forces teachers to "teach for the tests" and, therefore, neglect to give their pupils a well-rounded education. The countercharge is that well-made tests, if carefully chosen by the appropriate authorities, do reflect most of the important educational objectives that the teachers should be helping their students attain. But this latter argument fails to convince educational reformers, who contend that the tests nevertheless put the schools in such a straitjacket that new and better ideas about what and how to teach children do not get a chance to prove themselves.

3. Whether or not passing minimum competency tests should be required for promotion from grade to grade and/or for high school graduation: The argument for such a requirement runs in two directions. First, it is contended that in the absence of such minimum standards the high-school diploma has become a meaningless piece of paper; and second, it is contended that setting such standards will make more students work harder to master the skills they will need to get by in the adult world. The opponents have three comebacks: First, no one has yet been able to define "minimum competence" in a way that is any more meaningful than the usual high school diploma; second, the imposition of minimum standards will only encourage teachers to lower their sights for all the children; and third, attrition in the school population—especially among blacks and Hispanics—will skyrocket if the low scorers de-

cide that it is better to drop out early rather than to be branded as failures at school.

4. Whether or not multiple-choice tests* are detrimental to learning: Critics claim that such tests get at only the most superficial aspects of instruction, that many of the individual test items* are ambiguous, that they penalize creativity and critical thinking, and that they discourage the teaching of effective writing. Not so, say the defenders, who insist that a well-made multiple-choice test can often get at complex and sophisticated forms of thinking not ordinarily demonstrable in so-called free response tests;* that although a few ambiguous test items may escape the eye of the test maker, most are ordinarily caught and eliminated before a properly standardized test is put in final form; and that to make the tests a scapegoat for failure to teach good writing is only a cop-out for those who have yet to learn what good writing is or how to teach it.

5. Whether or not students' test results should be the basis for determining how well an individual classroom teacher is fulfilling the responsibilities of his or her job: This gets to the question of "accountability," which has stirred up so much bitterness among parents, teachers, school boards, principals, and politicians in recent years. The proponents argue that a student's low score on a standardized achievement test is evidence that the student's teacher is either incompetent or is willfully neglecting the specific educational tasks that the community has delegated to school teachers. The counterargument is that education should be a joint enterprise—in which the home and the community at large are also deeply involved—since the nature of learning is such that no teacher can be expected to do the whole job. In any case, it is argued, no one can tell from just looking at a test score what *caused* it to be high or low.

6. Whether or not organizations that conduct testing programs for admission to postsecondary educational institutions shall be required by law to publicize the questions and correct answers for every test immediately after its administration: One state (New York) enacted such a law in 1979, and the question whether the U.S. Congress should follow suit is, at this writing, being studied by the House Education and Labor Committee. Because this legislation is so recent, the pros and

cons of the debate have yet to be sorted out. In general, the proponents—led by student organizations and some legislators—argue that students need such detailed information (a) so that they can bone up to do better next time, and/or (b) so that they can have grounds for asserting that the test results are in error or have unfairly represented their true abilities. The opponents—mainly the testing organizations and the colleges—argue that if the specific questions and answers of any test are publicized, a brand-new set of questions will have to be constructed and analyzed each time the test is given. As a consequence, the cost of each test administration will skyrocket and the frequency of the testing will be greatly reduced, which will, they say, deprive many segments of the student population of the opportunity to take the tests.

Recently, representatives of five organizations responsible for major national testing programs announced a set of public interest principles to guide future developments in their standardized tests used for admissions to higher education. The guidelines include increased publication of the content of the tests, publicly visible procedures for eliminating test bias, and procedures to enable test takers to have their scores verified. They also call for procedures to increase the appropriate use of scores and to discourage misuse, for giving the test takers useful information about their performance on the tests and for giving them a voice in the design of the programs (see Appendix D). The test producers have invited reaction and discussion around their principles as the debate over national testing legislation continues.

The resolution of these and other issues related to testing in the schools is going to take time and a lot of clear-headed discussion among the parties concerned. Reaching such a resolution will depend largely on the degree to which the participants in the debate have achieved some common understanding of the basic principles of educational testing and measurement. The next chapter attempts to set the stage for this effort.

Educational Testing as Measurement

Some people shy away from the term "measurement" when used in connection with how pupils are doing in school. They feel that it implies more rigor than is desirable in the assessment of human learning. However, some understanding of how educational testing is used as a form of measurement is needed in order to make sound judgments about the quality of the testing your children experience and the dependability of the information it supplies. This chapter begins with a discussion of the elements of the testing process. It then describes the kinds of numbers, or measurements, ordinarily used to summarize the results of the testing. And finally, it looks at some hard questions about the reliability and validity of tests.

Elements of the Testing Process

A test can be thought of as nothing but the physical object, the batch of exercises printed in a test booklet. A test in the full sense, however, is a lot more than that. It is a rather complicated transaction between a test giver and a student, which is designed to rate the student's responses to a sample of tasks that have been specified by someone whom we have chosen to call the test maker (for example, a professional tester or a teacher). There are five steps in the testing process: (1) specifying the sample of tasks—e.g., the test questions or problems; (2) getting the students to respond to these tasks; (3) observing their responses; (4) rating the responses; and (5) interpreting the ratings. A glance at each of these steps may help you understand what testing is all about.

1. Specifying the tasks: In specifying the tasks, the test maker has many forms from which to choose. They may be tasks of the paper-and-pencil variety—essay questions, short-answer items, true-false items, multiple-choice items, etc. They

may involve the manipulation of objects such as a set of blocks or an instrument. They may be presented to the students orally or in print. They may be administered to a whole group of students all at once or to each student individually.

In working up the tasks, the test maker has to keep two things in mind. First, the task should be representative of all the tasks the students should be able to handle in a particular area. Second, each task must be framed in such a way that every student for whom the test is intended will see it the same way.

2. Getting the students to respond to the tasks:

This is central to the testing process, and it puts an enormous responsibility on the test giver, whether a teacher or someone else. It involves making sure in advance that every student knows what he or she is expected to do, creating an atmosphere that will make him or her *want* to cooperate, and then making sure that he or she does so. If this part of the testing act is not skillfully handled, the results of the testing will be useless.

3. Observing the students' responses to the tasks: There are essentially two ways in which the students' responses to the test may be observed. In the first way, the student's performance is observed while it is actually going on—for example, when it consists of giving a talk, playing an instrument, making bread. In the second way, which is probably the most common, some sort of record or product of the student's performance is observed—for example, the essay the student has written, the answers he or she has checked in a set of multiple-choice questions, the loaf of bread.

4. Rating the students' responses to the tasks:

This is the main purpose of testing: to provide an estimate of the overall quality of the responses each student has made.

When the observations are made while the student's performance is actually under way, the rater may either be the test giver or someone else specially designated to watch and judge. When the observations rely on some record or product of the performance, the rater may be the test giver or anyone else having access to the record—perhaps, in the case of multiple-choice tests, the "brain" in a test-scoring machine, which has been programmed by somebody who has been told by somebody else how to add up the answers.

The point of all this is that the rating of any student's performance is never more than an *estimate* based on the way the student has handled a sampling of tasks that someone has prepared for the purpose. In short, all educational testing, from first to last, is an intensely human transaction, with all that that implies about the uncertainties surrounding the appraisal of student performance.

5. Interpreting the ratings. This last step in the testing process is in many ways the most difficult and the most frequently neglected. It involves, first of all, deciding how much confidence you can have in the estimates of student performance the ratings provide. To what extent may they be overestimates or underestimates? How do they compare with other estimates of the same kind of performance as indicated by other tests, or by what the students' teachers have observed from day to day in the classroom?

Finally, the interpretation of any rating of your children's performance in a specific area of learning has to be examined in the context of their ratings in other areas of learning, and of the many factors that may be helping or hindering their educational progress. Without this sort of analysis, the time and effort that has gone into testing your children may not contribute very much to their education.

Translating Test Performance Into Numbers

The most familiar rating of student performance is a teacher's grade: "A" for excellent, "B" for very good, and so on down to "F" for failure. Despite the recent dissatisfaction with grading, there is little question that teachers' exam grades are an important way of showing you how your child is doing in school. The trouble is that, for any given test performance, one teacher's A might have been another teacher's C, D, or even F.

It was with the hope of helping schools around this problem that professional testers invented the so-called standardized test.* The hope—still pretty far from fulfillment—was that a test would become "standard" if it could be made to meet four requirements:

1. That the tasks in it would be, and would *seem*, the same to every student, wherever he or she goes to school.

2. That the testing conditions would be the same in all schools giving the test.

3. That the rating of every student's test performance would be done according to an agreed-upon set of rules.

4. That the ratings would be translated into a common set of numbers suitable to any test, whatever its content might be.

In working toward this last objective, testers usually have three kinds of numbers in mind, which they refer to as raw scores,* standard scores,* and percentile ranks.* These are admittedly technical terms, with which you may not be familiar, but it is important for you to have some understanding of their meaning in order to correctly interpret how your child has done on most standardized tests. Briefly, then, the significance of these terms is as follows:

Raw scores: These are the numbers that have come simply from adding up each student's answers to see how many have been judged to be correct. Thus, a math test with 24 problems to solve could have a range of raw scores from 0 to 24. A spelling test with 143 words for the student to spell could have a range from 0 to 143. An essay test with 2 questions, the answers to each of which could be rated 0 to 4, would have a possible raw score range of 0 to 8.

Standard scores: Since the possible range of raw scores can vary widely from one test to another, professional testers have tried to find a way of translating the raw scores of any test into a set of numbers (standard scores), which (a) will be the same for all tests, however long or short and whatever their subject matter, and (b) will provide comparable units of measurement. The methods by which the raw scores on any test are converted into standard scores are rather complicated, but roughly the steps go like this:

First, the test maker gives the test to a cross section of the kind of students for whom the test is eventually intended—for example, students in grades seven to nine, or twelfth-grade students who are applying for admission to college. Professional testers call this group of students the standard reference group*.

Second, the test maker then calculates the raw score for each student in this standard reference group.

Third, the test maker uses all these raw scores to get two statistics that summarize the performance of the standard reference group as a whole. One of these statistics is the average* of the group's raw scores; the other is a statistic (called the standard deviation*) that indicates how much the group's raw scores are spread out above and below the average.

Fourth, the test maker then uses these two statistics to translate the raw scores into a set of scaled scores*, which indicate how far above or below the average each particular raw score is.

The scaled scores so derived are called standard scores. The actual set of numbers in any standard score scale varies from one test publisher to another. Many publishers use a scale for all their tests that runs from 20 to 80. Others use a scale (called the stanine scale) that runs from 1 to 9. The standard score scale used for the College Board tests goes from 200 to 800. The standard score scale for most IQ tests usually goes from an IQ of 40 to an IQ of 160.

One thing to keep in mind with respect to any type of standard score scale is that the lowest score on the scale (e.g., 200 on the College Board scale) may not stand for the lowest possible raw score (zero), and the highest score on the scale (e.g., 800 on the College Board scale) may not stand for the highest possible raw score.

If standard scores seem a bit confusing, try re-reading the related material in Appendix B. And if that doesn't clear up the confusion, ask your local guidance counselor or testing director to explain in detail the meaning of the actual standard score your child received on some standardized test.

Percentile rank (PR): This is a number that is intended to help you see how your child's raw score or standard score on a test compares with that of all the other students who took the same test at the same time. It tells you what percentage of the students got raw scores at or below the score your child received. For example, suppose Mrs. Smith has a math class of twenty, seventh-grade students. She gives them the math test with twenty-four problems. Five students (i.e., 25 percent of her class) get raw scores of 9 or lower. This would mean that if Ramon got a raw score of 9, his percentile rank in Mrs. Smith's class would be 25 on that particular test.

But percentile ranks are a bit tricky. If the same

test had been given at the same time to all ten thousand seventh-grade students in Ramon's state, and if five thousand of them (i.e., 50 percent) got raw scores of 9 or lower, then Ramon's math test performance would give him a percentile rank of 50 *when his performance is compared with all seventh-grade students in the state as a whole*. Clearly, it is extremely important to know what group of students your child's percentile rank is based on.

It is also important not to confuse percentile rank with the percent of correct answers* your child makes on a test. Note that Ramon's raw score of 9 indicates that he correctly solved $37\frac{1}{2}$ percent of the math problems on the test, while the percentile ranks associated with that raw score tell you nothing whatever about the percentage of problems he solved correctly.

Reliability and Validity

In the jargon of professional testers, these two words, taken together, refer to the overall usefulness of a test. Therefore, it's important for you to understand what the testers have in mind when they use these terms. It is often said that validity* has to do with *what* a test measures, and reliability* has to do with the *accuracy* with which it measures whatever it is intended to measure. This is neat but a bit too simplistic. So let's take a closer look, first at the reliability question and then at the validity question.

The Reliability Question

The reliability question has two parts. First of all, you need to get some idea of how careful the raters of test performances are likely to be in arriving at the numbers by which the ratings are expressed. This component we'll call the *accuracy of the rater*. Second, you need to consider how adequate a sampling of any student's responses the testing may have provided. This aspect of the reliability question we'll call the *adequacy of the sample*. Some hypothetical examples may help to clarify the distinction.

Accuracy of the rater: The job of grading a batch of answer sheets containing the responses students have given to, say, two hundred true-false questions illustrates in the most elementary way what has to be done to determine the accuracy of

the rater. A system of checks has to be set up to see how often the rater—whether a clerk or a test-scoring machine—may have miscounted. This may seem trivial, but both clerks and machines have been known to err, so that 100 percent accuracy can never be absolutely guaranteed. A well-organized testing operation will make spot checks periodically to see how rater accuracy is holding up.

A bigger problem is the kind of rater accuracy encountered in the grading of essay tests. Here, one rater's judgment of what a student has produced may differ markedly from that of another rater with different background and tastes. Consider the case of Brown and Jones, two English teachers who have worried about the accuracy of their grading of student themes. They pick at random ten themes written by Brown's students and ten written by Jones's, and remove all clues from the papers that might identify who the writers are. Then, working independently, each teacher grades all twenty themes in the usual manner, A to F.

If Brown and Jones agreed in the grades they assigned to only 20 percent of the papers, the rater accuracy (examiner reliability*) would be considered extremely poor—no better than what would be expected if the two teachers had assigned the grades at random without reading the themes. If, however, their grades on fifteen of the twenty papers were in agreement (75 percent agreement), the rater accuracy would be considered exceptionally high. If, though, the inter-rater agreement were 100 percent, you might expect that there had been some skulduggery: maybe Brown sneaked a look at Jones's grades and copied them, or vice versa.

Adequacy of the sample: You need to remember that any test can never be more than a *sampling* of your child's performance. It is two kinds of a sample. It is a sample of the multitude of tasks in some clearly specified field of learning. It is also a sample of all the responses your child might have made to each particular task under different testing conditions. It is, therefore, important for you to have an estimate of how much your child's test results would vary over a large number of such samplings. This estimate is sometimes referred to as the *sampling error**, or, more commonly, as the *standard error of measurement**. It encompasses

all the sources of error in arriving at any test result; it is, therefore, the most important part of the reliability question, for the bigger the standard error of measurement, the lower the reliability of the test.

How is the estimate of the error due to sampling made? One way consists of making up two tests. Test A would consist of one set of tasks, and Test B would consist of an equal number of different tasks, which would nevertheless cover the same kinds of subject matter as in Test A. The two tests would then be given to an appropriate group of students on successive days to see how each student's score on Test B might differ from his or her score on Test A. The size and direction of these differences could be expected to vary in some amount from student to student.

Now suppose, for example, this exercise had been carried out with alternate forms (Test A and Test B) of the twenty-four-problem math test mentioned above. On a well-constructed, carefully scored, and properly administered test with this number of problems, Mrs. Smith would most likely find that the scores on Test B of two-thirds of the students will differ from their scores on Test A by no more than plus or minus 2 points. Then, ± 2 would be the estimated *standard error of measurement of this test*. Thus, if Ramon got a score of 9 on Test A, the odds would be two-to-one that his score on Test B would be no higher than 11 and no lower than 7. In terms of percentile rank, this would mean that Ramon's parents should not be too surprised if his math performance on Test B ranked him as high as the 50th percentile or as low as the 20th percentile in Mrs. Smith's class.

This wobbliness due to sampling is a fact of life in all educational testing, whether the tests be teacher-made or standardized tests. A more extensive sample can reduce the wobbliness somewhat, but it can never be wholly eliminated. So when interpreting your child's test score, you had better pay close attention to the estimated range of sampling error, which the test maker should provide. (For further discussion of the standard error of measurement see Appendix B.)

The Validity Question

The validity of a test has to do with its relevance to some specific purpose which the test user has in mind. Professional testers and psychologists

1

speak of so many varieties of validity (content and construct validity*; internal, external, and concurrent validity; criterion-related and predictive validity) that it is hard for test users—and even some testers—to distinguish among them. There are two kinds, however, that are particularly relevant to educational testing—content validity* and predictive validity.* These two kinds raise questions about the quality of the local testing process, questions to which test users themselves must be largely responsible for finding the answers. This will become apparent as we take a closer look at these two kinds of validity and at the things that may threaten the validity of the testing that goes on in any school.

Content validity: Before your school system purchases a battery* of achievement tests, it should first question whether the content of each test is in line with its own curriculum. Do the questions in the science test intended for junior-high students reflect what is being taught in your junior-high science class? Are the problems in the math test for elementary schools representative of the kind of math being taught in your elementary school?

Questions of this sort are questions about content validity. The answers to them can be gotten in a straightforward way by going over the tests question by question to see what skills are supposed to be covered. If the publisher's junior-high science test contains a lot of questions on meteorology and only one or two on biology, the test would hardly yield valid information about students' science achievement in a school system that puts major emphasis on biology in the eighth grade.

The question of content validity applies to teachers' homemade tests as well. For example, suppose the local curriculum committee has decided that an important objective of elementary math instruction is to get students to handle problems that they will face as consumers, such as:

A 20-ounce jar of grape jelly is priced at \$1.20. An 8-ounce jar of the same jelly is priced at 64 cents. In price-per-ounce, what is the difference between the 20-ounce jar and the 8-ounce jar?

Now suppose the teachers in some of the schools never get beyond testing their own pupils with problems like this:

$$1.20 + 20 = ?$$

$$.64 + 8 = ?$$

In light of the agreed-upon objectives in math, the tests of those teachers are lacking in content validity, for they have tested only for basic computing skills and have not tested to see whether students *can apply* those skills to the type of problem they might have to solve in any supermarket.

Predictive validity: In this case the validity question is: How helpful might a test score, or a series of test scores, be in forecasting what a student's performance will be a month from now or a year from now? It is the sort of question asked to help decide such matters as whether a student would be better off in one program of study than in another, or whether the student has a better chance of success in College A than in College B. The answer to this type of question can never be exact, and it can change as circumstances change. Nevertheless, it is a question you must face if the decisions about your child's future are to rest on anything more than a random hunch.

Take, for example, the use of the College Board's Scholastic Aptitude Test (the SAT) for admission to College A. What College A needs is an estimate of each applicant's chances of making a satisfactory record in the freshman year. As one basis for getting such estimates, it runs a so-called prediction study on the records of students admitted in the preceding year. From an analysis of the data, it can get a fairly good idea of how well the SAT scores can forecast the grades of the next group of applicants. A somewhat oversimplified description of the results of such a study might look something like this for a class of two hundred students, twenty of whom ended up the freshman year with a failing grade-point average:

1. Of the 180 students with passing grades, 98 percent of them entered with SAT scores of 500 or above; 2 percent entered with scores below 500.¹

2. Of the twenty students with failing grades, 20 percent entered with scores of 500 or above; 80 percent entered with scores below 500.

¹500 is a fixed point that marks the middle of the College Board's standard score scale. It should *not* be regarded as an average. Its meaning in terms of level of student performance has remained the same over at least the last twenty years. It does not fluctuate in accordance with changes in the group of students that happen to be taking any particular version of the test in any given year.

This outcome tells College A that, as things stand now, an applicant with an SAT score of 500 or above may have a 98 percent chance of making the grade in the freshman year at this college, whereas an applicant with a score below 500 may have only a 20 percent chance. A result like this would indicate that, for the time being, College A is getting just about the most possible predictive power that the SAT can provide. Another college might well find that the scores were far less useful in forecasting the probable success of its applicants, since a test may show good predictive validity for some students in some places and poor predictive validity for other students in other places. Thus, it doesn't make much sense to talk of *the* predictive validity of any test.

Threats to validity: The validity of your child's test results can be undermined in various ways. Perhaps the biggest threat comes from *the passage of time*. Like everything else in the world, schools and the people in them are constantly changing. A test that was validated last year should work almost as well this year. But if it was validated five years ago, the test may not be very useful today.

Cheating is, of course, another obvious threat to a test's validity. Clearly, if a student copies another student's answers, or gets another student to take the test in his or her place, the validity of that student's test results is nil.

Improper coaching can also upset the validity of a test. There are two kinds of coaching: proper and improper. Proper coaching aims to get students thoroughly familiar, through advance practice, with the mechanics of the test-taking process, with the types of tasks the test contains, and with the best strategies for approaching each type. Improper coaching is the same thing as cheating. A teacher, or someone else, gets a copy of the test in advance and uses it to drill the students until they have memorized all the answers, whether or not they understand *why* any answer is correct. This practice may make the students' test results look good, but it obviously destroys any validity the test might have had for measuring anything other than the students' ability to memorize.

Poor test administration can be another serious threat to the validity of a test. When students are confused by the instructions for taking a test, and when the test givers are careless about such mat-

ters as the adequacy of the seating arrangements, the creation of a congenial atmosphere in the examination room, the accuracy of the timing, and other particulars, the results of the testing will be something less than useful.

Low reliability is, of course, also a threat to validity. If the scores on a test are afflicted with a huge standard error of measurement, the validity of the test is certain to be low or nonexistent. On the other hand—and this is a point you should always keep in mind—even though a test may have very high reliability (i.e., a very small standard error of measurement), the test may be quite lacking in validity for the particular purpose a school has in mind. For example, a highly reliable spelling test may have little or no validity as a measure of creative writing.

Pitfalls in Communication

Any discussion about educational testing is likely to be confusing, because many of the key words in the testing jargon may mean different things to different people. As we have shown in the preceding chapter, the word test itself can be thought of as nothing more than the exercises printed in a test booklet, or as all the steps involved in the complex testing process. In this chapter, we will take a look at some other terms and concepts that commonly cause difficulties in communicating about educational tests and test results: norms*, grade equivalent scores*, mental age*, IQ and aptitude, over- and under-achievement, and minimum competence.

Norms

Most publishers of standardized achievement and intelligence tests always provide a table of norms to help interpret the test results. This is why some test makers prefer to call the standardized variety "norm-referenced tests*."

For achievement tests, the norms are usually reported in two sets of tables. One set shows the percentile ranks of the scores made by individual students at each school grade; the other shows the percentile ranks of the score *averages* of whole school systems, also for each grade. Both sets have ordinarily been obtained by testing students in a sample of schools, who are supposedly representative of all the students in all the schools in the nation. Thus, after your child, say an eighth grader, has taken a standardized test in reading designed for junior-high students, you can consult the appropriate norms table to find out how your child's reading performance compares with the reading performance of other eighth graders who have taken the same test under similar conditions.

Let's say that, after consulting the norms table, you find a number indicating that your eighth-grade student's reading performance is at the 20th percentile of eighth graders in general. You should keep in mind that your student's percentile rank could have been different if:

- he or she had taken an alternate form of the same standardized test;
- he or she had taken a standardized test in the same subject but produced by a different test maker, who had different ideas about the kinds of exercises that constitute reading ability;
- he or she had taken the test at a different time in the school year than when the norming operation had been carried out;
- the norms data had been derived from a different sample of schools with a different mix of students;
- the organization of instruction in your student's school is different from that in the schools involved in the norming operation;
- the conditions for testing in your student's school were more favorable or less favorable than those in the schools that cooperated in the norming operation;
- the norms data had been gathered some years ago when the scores from which the percentile ranks were derived may have run generally higher or lower than they do today.

In short, the norms accompanying any standardized achievement test are not absolutes. Nor do they provide the basis for establishing fixed and invariable "standards" of academic performance. It is easy to forget this for two reasons. First, because the word "standardized" may give the impression that the gathering of norms is synonymous with the setting of standards of good as opposed to poor performance. Second, because historically the word norm has indeed been commonly used to mean such things as "a principle of right conduct; an ideal standard binding upon the members of a group and serving to guide, control, or regulate proper and acceptable behavior" (*Webster's Third New International Dictionary*). In the case of an achievement test norm, however, we are dealing not with "right conduct" or "ac-

ceptable behavior," but with a *number*—namely, the test score that happens to be equivalent to a percentile rank of 50 for the particular sample of students who participated in the norming operation. To what extent a number should be used to define an "ideal standard" of any kind for any school or student is a question that needs to be deeply pondered.

Because conscientious test publishers are aware of the ambiguities inherent in national norms, they often urge test users to work out "local norms"—that is, the percentile ranks, grade by grade, based on the raw scores of *all* the children in *all* the schools of the state or the county or the local school district. This procedure has two main advantages: (1) It does away with the question of whether the students in the norming sample are adequately representative of all children, or whether such a widely representative sample is relevant to the students in your school; (2) It makes it more likely that the tests used have been selected by the local authorities because their content is valid for the local curriculum.

Grade Equivalent Scores

A commonly heard statement goes something like this: "The test shows that your eighth-grade child is reading at the fifth-grade level." Or like this: "The test shows that your seventh-grade child is reading at the twelfth-grade level. Wow!" Or like this: "The tests show that the average reading performance of the sixth graders in this school is at the third-grade level." Statements like these are unfortunate because they usually reflect an unquestioning acceptance of grade equivalent scores as measures of academic achievement. In addition, such statements don't take into account the fact that schools may vary widely in the ways they sequence the instruction they offer.

Grade equivalent scores came on the educational testing scene a little over fifty years ago. The hope of the inventors of grade equivalent scores was (1) that they would form a single score scale that would satisfactorily encompass the whole range of student achievement from kindergarten through grade twelve, and (2) that such scores could be readily and correctly interpreted by test users who might be unfamiliar with the niceties of educational measurement. However,

no score scale of such extended scope has yet been found to be technically satisfactory. And the use of grade equivalent score scales in particular has fostered some weird misinterpretations of what the results of educational tests mean. This is the reason that, during the last twenty years or so, more and more test makers have been urging that the use of grade equivalents be abandoned altogether.

Why should grade equivalent scores be regarded with misgivings by many of the test publishers who provide them? It is because they know that the great majority of the grade equivalent scores are based on norms data that is either shaky or nonexistent.

Suppose a math test with forty problems has been normed nationally by giving it in the fourth month of the school year to a national sample of students in grades six, seven, and eight. For this sample of students, the average raw scores and their grade equivalents would look like this:

Average raw score	Grade equivalent score
15	6-4 (6th grade, 4th month)
18	7-4 (7th grade, 4th month)
23	8-4 (8th grade, 4th month)

Nothing is known for sure about what the average raw scores would have been if the students had been tested either earlier or later in the school year, or in grades below the sixth and above the eighth. Yet the table of grade equivalents that the test maker supplies with the test gives a list of scores purporting to represent what the average raw scores would have been had the test been given to the students in, let's say, the second month of the third grade or in the fourth month of the eleventh grade. For example, the table might tell you that a raw score of 38 on the forty-problem math test would be the average for all students in the seventh month of the twelfth grade ($GE = 12 - 7$). Thus, if your seventh-grade child comes home with a GE of $12 - 7$, you are asked to believe that his or her math performance is at the same level as that of students getting ready to graduate from high school, *even though the test may contain no math problems ordinarily dealt with in grades nine through twelve.*

Furthermore, if you take grade equivalent scores at face value you might come up with some

rather extraordinary conclusions. For example, you might infer that:

- Practically every student in every grade nationwide must be above or below grade level: only the few students whose raw scores happen to be exactly at the average for their grade and month would be regarded as performing "at grade level."
- If your eighth-grade child temporarily loses his or her place in the math test booklet and thereby accidentally misses four problems, the chances are that he or she would have lost a whole year in grade equivalents.

The moral of this strange tale is: Do not put your trust in grade equivalent scores; always ask to see the tables of raw scores from which they have been derived.

Mental Age (MA)

In 1908, the French psychologist Alfred Binet invented mental age as a measure of intelligence and thereby laid the groundwork for a vast amount of confusion and controversy, which has been going on ever since.

Mental age scores have most of the same characteristics as grade equivalent scores and can lead to similar gross misinterpretations. In fact, the mental age scale could just as well have been called an "age equivalent scale." This means that you can come up with the same kinds of absurdities when you try to translate the raw scores on an intelligence test into years and months of age.

Because of this, most publishers of intelligence tests these days have abandoned the mental age concept altogether. So, when someone says gloomily, "Your twelve-year-old child has a mental age of ten," or, more cheerfully, "Your ten-year-old child has a mental age of fifteen," that person may be seriously out of touch with accepted practice in the interpretation of children's performance on intelligence tests.

IQ and Aptitude

The demise of the mental age concept has not affected the IQ's strength as a source of confusion among test users. As you probably know, the abbreviation "IQ" stands for "intelligence quotient,"

which in the past has been computed by dividing mental age by chronological age (MA/CA) and multiplying by 100 to get a round number.

But in point of fact, the IQs associated with most modern intelligence tests are no longer calculated by the above method and are not "quotients" in any sense of the word. They are standard scores that form a scale running from a low of about 40 to a high of about 160. Test makers refer to the old, now discredited quotient type IQ as the "Ratio IQ*" and to the new standard score type IQ as the "Deviation IQ.***" (See Appendix B for a more detailed description of how Deviation IQs are calculated.)

It has occasionally been suggested that schools, pupils, and parents would be better off if the term IQ were expunged forever from the vocabulary of testing. Too many people forget—or have never realized—that the IQ, like any test score, is *nothing more than an index number that is useful for summarizing how a student has responded to a miscellaneous series of test questions.*

Since IQ is just a test score, what does it mean when someone says to you, "This test measures your child's IQ"? Literally, of course, it means, "This test measures your child's score on an intelligence test!" But most people use the term IQ as shorthand for something they think of as "intelligence" or "mental ability" or "aptitude." Because of this, IQ tests have become the center of controversy about the nature of mental ability. There are three divergent opinions about this, which go as follows:

1. Mental ability is some sort of inherited, unchangeable entity that *completely* determines for all time how high or low your child's IQ score will be and how well or poorly your child "ought" to do in school.

2. Though mental ability is indeed some sort of inborn entity, it only *partly* determines the level of your child's IQ score, because that score is also in part determined by the child's environment, especially his or her early environment.

3. The idea that mental ability is some sort of inborn entity is just a lot of poppycock, for everybody knows that your child's IQ score is *wholly* determined by the child's environment at home and at school.

These three views are in line with the unresolved debates that have been going on among psychologists and others for at least the last sixty years concerning the roles of heredity and environment in human development. But there is a fundamental flaw in all three lines of reasoning. *They all rest on the impossible assumption that anyone, simply by looking at your child's IQ (or mental ability or aptitude) score, can tell what caused the responses it summarizes.* As Dr. Anne Anastasi, a widely recognized authority on psychological testing, has so wisely and cogently pointed out:

Tests are designed to show what an individual can do at a given point in time. They cannot tell us *why* he performs as he does.

In a word, a student's test score, whether it be an IQ score, an aptitude test score or an achievement test score, can be educationally helpful if the information it provides can be used to figure out *right now* what can be done to help the student improve his or her performance in the future. It is useless if it is merely used to explain that the student's high or low test performance is caused by some unobservable "inborn entity," or some vague concept called "social environment."

It has been shown that by giving a student appropriate guided practice in the cognitive skills sampled by IQ tests, a student's IQ can be raised just as a student's score on a spelling test can be raised by teaching him or her how to spell. Some students may need more of the guided practice over a longer period of time than other students need, and the kinds of practice needed can differ from student to student. One of the important benefits of a well-designed IQ test is that, in the hands of a perceptive test giver, it can provide useful clues about the specific kinds of guided practice your child needs to improve his or her cognitive skills.

Over- and Under-Achievement

Another popular misconception, which grows out of the notion that the IQ is a measure of something fixed in a child at birth or in the years of early development, is what might be called the over- and under-achievement fallacy. It leads to such results as the following:

1. If your child's IQ score is at the median* (the 50th percentile) of his or her age group (IQ = 100),

and if your child's overall score on a battery of achievement tests puts him or her at the 25th percentile according to national grade norms, a careless user of test results is likely to label your child an "under-achiever," or as "a pupil who is performing below capacity."

2. If your child's IQ score is at the 25th percentile of his or her age group (IQ = 90), and if your child's overall score on a battery of achievement tests puts him or her at the median (the 50th percentile) according to national grade norms, a careless user of test results is likely to label your child as "over-achiever," or as "a pupil who is performing above capacity."

There are three things you should understand about the relationship between IQ testing and standardized achievement testing. First, both types of tests are likely to be made up of so many of the same kinds of questions that if the test titles were removed it would be hard to tell which was which. Second, the aim of both types of tests is to get a measure of what children have *learned*. And third, the difference, if any, between the percentile rank of a pupil's IQ score and the percentile rank of the same pupil's total score on a battery of achievement tests is largely, if not wholly, chance. The difference may be solely the result of errors of measurement in the two test scores. It may be the result of differences in the manner by which the norms for each type of test have been obtained. It may be the result of differences in the methods by which each test was administered.

The most that can legitimately be said about any difference between a student's percentile rank on an IQ test and his or her percentile rank on a battery of achievement tests is that the score on the first has either *underestimated* or *overestimated* the score on the second. Such discrepancies are just another indication of the uncertainties in the appraisal and prediction of pupil performance, so be wary of labels like "over-achiever" and "under-achiever."

Minimum Competence

Many of the same problems we discussed in connection with norms, grade equivalents, and IQs apply to the use of tests for distinguishing between competence and incompetence in the so-called

basic skills*. Minimum competence is sometimes defined as an IQ score of 70, which translates to a raw score that is just slightly above the second percentile. Some intelligence testers feel that the minimum IQ should be higher, say an IQ of 80, so that all the children who really cannot make it in school are screened out. The percentile rank of an IQ of 80 is about nine in most norming samples.

Grade equivalent scores have also been suggested as a way of defining minimum competence. It has been seriously proposed, for instance, that if a child winds up the school year with achievement scores below the national median for his or her grade, then that child should be "retained" in that grade until he or she has made it to the median (in which case half the children would be held back!). In other instances, it is being proposed that in order to receive a high-school diploma the student must, at a minimum, have demonstrated the ability to read, write, and figure "at the ninth-grade level."

This tendency to conceive of minimum competence in terms of test norms has worried test makers recently. For this reason, a number of test makers are now leaning toward criterion-referenced tests as a more promising approach to the definition of minimum competence. These tests focus on the kinds of concrete tasks that presumably everybody has to cope with in order to be minimally self-sufficient citizens—tasks such as:

- Can the student read traffic signs well enough to know what the speed limit is?
- Can the student fill out an application for a job well enough to be understood by a prospective employer?
- Can the student figure the relative cost of making a purchase on an installment plan as compared with paying cash at the time of purchase?
- Does the student know what a primary election is for and what he or she must do in order to vote in one?
- When faced with a health emergency, does the student know how to call for help and what to do until an ambulance or first aid arrives?

This approach was largely stimulated by the work that went into the National Assessment of Educational Progress, which got under way in the

1960s and has been under the sponsorship of the Education Commission of the States. The same approach has since been taken up by a number of state education agencies. It has opened up all sorts of new problems for test makers and test users, not the least of which are those that revolve around the extraordinarily difficult question: What person or organization decides who decides which specific tasks every student must successfully perform to demonstrate that he or she is minimally functional?

The jury is still out on this question. But the fact that it is being asked and earnestly debated in some quarters is a hopeful sign that people are becoming less willing to hide behind the magic of numbers in discussions of how schools and the students in them are doing.

CHAPTER 4

Policy Matters

In many respects, the politics of education and the politics of educational testing come to much the same thing. This is apparent in the various ways that special interest groups use or attack test results to advance their divergent views about what schools should be doing for children. There are three overlapping problems in particular that have recently become implicated in the politics of education: bias in testing*, testing for educational accountability, and testing for minimum competence.

Bias in Testing

How do you decide whether a test is biased for or against any group of students? Probably the usual approach is to use your intuition: you simply examine the test questions and sense that too many of them would be unfair or inappropriate if used with certain groups of examinees. For example, a reading test loaded with material about the experiences of middle-class white children spending a summer vacation in the country might be regarded as inherently biased against poor black and Hispanic children who spend their summer in the city streets.

The intuitive approach is useful up to a point, but it doesn't work if the people making the hunches have biases themselves. To deal with this problem, test researchers have been trying to get evidence on test bias by analyzing test results. In essence, what they are after is an explanation of why a given test may have less reliability and/or less validity for some groups of test takers than for others. The three hypothetical cases that follow are intended to give a rough idea of the types of evidence the testers are looking for.

Case I: Evidence of examiner bias. In Chapter 3, we discussed inter-rater reliability by comparing the grades that Brown and Jones might have given to the same set of twenty student themes. Let's

now assume that the result of that hypothetical exercise showed Brown and Jones in good agreement. It should be recalled that one of the stipulations for estimating inter-rater reliability in this instance was that neither examiner was to know the identity of any student whose theme was being graded. Let's now get Brown and Jones to grade another set of themes written by the same twenty students. This time, though, we'll change the rules a bit: (1) Jones is to know the identity of each student whose theme he or she is grading, while Brown is not given this information, and (2) by virtue of knowing who the theme writers are, Jones also knows that ten of them are girls and the remaining ten are boys.

Under these changed conditions, the amount of agreement between Brown's and Jones's theme grades could suffer a sharp decline. This time we find Jones giving the girls' themes nothing but A's and B's, while Brown (who is unaware of the sex of any of the theme writers) gives approximately the same proportions of A's, B's, C's, D's, and F's to both the girls' and the boys' themes. This could be taken as evidence that Jones biased the grades in favor of girls and against boys.

Of course, this would not be conclusive evidence that Jones's judgment of theme quality was warped by foreknowledge of who the theme writers were. But if Jones were to admit that he or she had approached the task of grading those themes while being deeply convinced beforehand that girls as a group *a/ways* write better than boys, the evidence for examiner bias in Jones's case would be pretty strong.

Case II: Evidence of bias in objective achievement tests. Suppose that the third-grade pupils in all the schools of a large city were given a fifty-item test in reading. Suppose further that the average raw score made by the pupils in School A is 39 and the average raw score made by the pupils in School Z is 21. An analysis of the school records shows that 80 percent of the children in School A come from English-speaking families with incomes over \$15,000 a year, while 80 percent of the children in School Z come from Spanish-speaking families with incomes under \$15,000 a year.

Can the 18-point difference in the raw score averages be regarded as evidence that the reading test has a built-in bias against poor children from homes where little or no English is spoken?

Some people would say "Yes!" unequivocally on the ground that it is unfair to test children from Spanish-speaking homes with a reading test in English—they should be tested in the language with which they are most familiar.

On the other side, there would be people—including some of the Hispanic parents themselves—who would reject this argument as missing the point of what achievement testing in the schools is all about. They would point out that, since one of the major objectives of American schools is to teach children to read standard English, the function of an achievement test in reading is to find out how well the pupils have learned to read standard English, not Spanish or any other language. A test should not be called biased on the ground that it is not measuring what it was never intended to measure.

At this point, the argument could go in several directions, leading to some complex questions about social policy and the rights of children. It could be argued, for instance, that although the reading test may not have a built-in bias, its indiscriminate use in the schools reflects bias in the educational *system* for not recognizing that American society is multilingual and culturally diverse. Who has the right to say that *all* American children, whatever their roots, are required to learn standard English and be tested in it? And the counterargument becomes: Who has the right to *deny* any child the opportunity to learn standard English and be tested in it?

These are large issues that go well beyond the merits or defects of the tests per se. In the heat of controversy, there is, however, an important point that may be overlooked. All sides might theoretically agree that, if carefully developed, the *test instrument* in itself can be free of any built-in bias. However, it's possible, in certain cases, that the *overall test situation* may cause certain groups of children to react in ways that tend to drag down their test performance. For example, it might have been observed that while the testing in School A went off without a hitch, in School Z the pupils were so turned off by the whole business that many of them refused to make any effort at all, and created so much disturbance in the examination room that the rest of their classmates were prevented from doing as well as they might otherwise have done.

Case III: Evidence of bias in a test for predicting academic success. A lot of attention has been given recently to figuring out what kinds of data would indicate whether an aptitude test is biased against a minority group seeking admission to college. This effort has run into difficulty for the reason that, although aptitude tests have long been used as aids in predicting college grade-point averages, the evidence of aptitude test bias available from any institution is likely to be ambiguous. However, there is pretty good agreement on what the results from a biased aptitude test *would* look like if colleges could produce the data.

If, for example, the black students admitted to College X made the same percentage of satisfactory grades in their freshman year as the white students, even though their average aptitude score was considerably lower, the aptitude test could be considered biased against the blacks.

Testing for Educational Accountability

The idea of accountability as applied to schools and school systems is nothing new. As far back as 1647, the government of the Massachusetts Bay Colony passed a law requiring that all the children in the colony be taught to read and write in order to foil the designs of the devil on their souls. The governors held each town *accountable* for providing such instruction, and they enforced the mandate by levying a fine of five pounds on every town that failed to comply.

Nowadays, each state has the same sort of power over local educational operations. However, the degree to which the power is exercised and the manner in which it is exercised vary enormously from one state to another. Much depends on the moods of legislators, the amount of clout the state education department exerts, and the quality of the information on student progress that is available from the school districts. To keep abreast of the developments in testing for educational accountability, the Education Commission of the States makes periodic surveys of what is happening in each state (see Appendix A). These reports can be of inestimable value if you are trying to find out how accountability testing, and its sibling minimum competency testing, are faring in your state, as well as in the nation at large.

In essence, this kind of testing, whether it be

under state control or under local control, seeks answers to two worrisome questions: (1) How much and what kinds of student learning is the educational tax dollar buying?; (2) Who is to be held accountable (i.e., who is to get the *blame*) if the average scores for any school are below those for the district, the state, or the nation?

These are threatening questions. They carry the implication that, if the test results appear unfavorable, school funds will be cut and/or somebody will be punished. Such fears are not groundless. For example, not long ago a state legislature considered, though it did not finally enact, a bill that read in part as follows:

If the performance of any school district on any test approved by the state board of education . . . does not equal or exceed the national performance average for such a test for two successive years, said school district shall not receive any further state financial assistance . . . until such time as said school district has achieved such national performance average.

It is the atmosphere created by threats like this that has led some state and national teacher organizations to call for the abolition of norm-referenced testing altogether. And where such testing has been mandated, you'll come across various kinds of defensive behavior that defeat any educational purpose the testing might have had. This can happen, for instance, when the effectiveness of a school is to be judged by the score gains its students make when one form of a test is given at the beginning of the school year and an alternate form at the end of the year. Such gains have occasionally been inflated by shortening the time limits on the tests administered in the fall and lengthening them on the tests administered in the spring.

In some districts where the test results, school by school, are published annually in the newspapers, some teachers have resorted to the questionable practice we mentioned previously of getting hold of the actual tests in advance and drilling their students until they memorize all the correct answers, even if they don't know why the answers are correct. Such a practice, of course, destroys any validity the results might otherwise have had.

When you come across this kind of drilling, you need to be careful about questioning the motives of those who engage in it. No doubt it is done by some teachers who, in all innocence, fail to distinguish between a test as an instrument for measuring educational progress and a textbook exercise for bringing about such progress. There is, furthermore, a line between legitimate and illegitimate "coaching" for tests that may not be generally recognized.

For example, in one sizable city the director of testing, who sought to obtain unbiased comparisons among the score averages of his elementary schools, wanted to make sure, insofar as possible, that every child in every school would be equally familiar with the various types of questions on the tests he had decided to administer. He didn't think that the practice exercises printed in the test booklets themselves were adequate for this purpose. He, therefore, prepared a considerable quantity of such practice exercises and asked the teachers to use them to instruct their pupils in the mechanics of the test-taking process and in the most effective way of attacking each type of question. The goal was to eliminate any bias in the test results that could arise from having pupils in some schools come to the actual testing with less knowledge about how to approach the test than those in other schools.

This is a sensible and legitimate form of coaching, aimed at minimizing a type of test bias that, if not corrected in this way, could easily invalidate the results from certain schools. However, when the press and the public heard what was going on, the testing director and the teachers were lambasted for trying to make some of the schools look better than they really were.

Testing for Minimum Competence

Until recently, accountability testing has relied almost exclusively on norm-referenced tests. Within the last ten years, however, many states and school districts have been turning to criterion-referenced tests as presumably a more reliable way of checking whether students and schools are succeeding or failing to do the job expected of them. This development has led to some new problems, which the test experts themselves, as well as the public and the school personnel, are having

trouble with. Part of the reason for the problems is that the development of criterion-referenced tests for measuring minimum competence is still at an early stage.

Another factor that makes the question of minimum competency hard to resolve is the failure on the part of some people to differentiate among the ways in which criterion-referenced tests are being used to identify and to deal with minimum competence.

There are four such uses that can be recognized as fairly distinct from one another as educational strategies:

1. For monitoring the teaching-learning process as it goes on hour by hour in the classroom.
2. For deciding which students should be assigned to special classes for remedial instruction.
3. For deciding at the end of the school year whether a student should be retained in the same grade for another year.
4. For deciding at the end of the twelfth grade whether a student should be denied a high school diploma.

Below is a sketch of each of these practices, along with some of the controversial points of each.

For hour-by-hour use in the classroom: It is in connection with this use that the modern concept of criterion-referenced testing first came into use. It represents a testing-teaching strategy that has still not penetrated many school districts.

The aim is to provide the teacher with a way of keeping track of how well each student is mastering the basic skills, so that instruction can be continually geared to the student's learning needs. Let's say, for example, that one objective of a fourth-grade curriculum is to get every pupil competent enough in the use of the alphabet that he or she is proficient in looking up unfamiliar words in the dictionary. The first step would be to see which, if any, of the pupils had not yet mastered their ABC's at the most elementary level. For this purpose, the teacher might begin by giving the class a five-minute, paper-and-pencil test consisting of items like these:

1. What letter comes before M?
2. What letter comes after V?
3. What letter comes between F and H?
4. What letter comes after T?

If a child were to get fewer than 80 percent of such items right, the teacher might decide that the child's knowledge of alphabetical order is below the minimum needed to go on to exercises of greater difficulty. So the child would be given more work on the ABC's until he or she had reached the 80 percent level on a retest like the one above. Having achieved this level of competence, the child would then be given some tasks to familiarize him or her with a standard dictionary, and then given another short test consisting of items like these:

Where would you find each of these words in the dictionary? Underline the right answer.

1. come: (a) near the beginning (b) near the end.
2. take: (a) near the beginning (b) near the end.
3. ask: (a) near the beginning (b) near the end.
4. you: (a) near the beginning (b) near the end.

Again, pupils below the 80 percent cut-off score would be judged less than adequately competent at this level and would be given more of the same kind of dictionary work until they could demonstrate enough mastery. The same step-by-short-step approach would be continued with each student until all had reached the 80 percent level on a test that might look something like this:

In each group of three words, underline the *one* word that you would find nearest the *end* of the dictionary.

1. laconic—laboratory—latent
2. minimum—misanthrope—microbe
3. reliant—relax—regulate
4. correlate—cormorant—corolla

Pupils would need different amounts of instruction to get to the point where they could pass this last test, but the hope is that they would all get there in a reasonably short time.

This step-by-short-step approach makes testing for minimum competence an integral part of the strategy for individualizing instruction in any of the basic skills. The principal arguments in favor of this procedure are:

1. It pinpoints for students and teachers the kinds of competence in the basic skills that are truly "basic" to further learning.

2. It takes account of individual differences in the rates at which students acquire any particular skill.

3. It gives each student as many opportunities as he or she needs to master each skill.

The principal arguments against the procedure are:

1. It assumes without adequate proof that the skills being taught and tested for are indeed truly "basic" to each student's further learning.

2. It overemphasizes purely academic skills and leaves little opportunity for the teacher to encourage the kind of social and emotional development of students that can come only through participation in group activities.

3. It makes the unlikely assumption that the kinds of skills apparently needed for coping with the world of today will be the same as those that students will need for coping with the world they will face as adults twenty years from now.

Whatever you think of the arguments pro and con, the decision whether or not to use this step-by-step approach to minimum competency testing probably rests on how much it will cost the taxpayers of the local district. Such costs can be considerable, since they include those for new bundles of tests and the teaching materials to go along with them, as well as the costs for training classroom teachers who have been accustomed to other ways of teaching the basic skills.

For assigning pupils to special classes for remedial instruction: This involves using criterion-referenced tests at the beginning of each school year and using the cut-off scores to sort out pupils who are so far behind their classmates in any basic skill that they require special instruction to catch up.

The principal arguments in favor of this approach are as follows:

1. It permits the regular classroom teacher to move ahead with the majority of the students as a single group.

2. It ensures that pupils below the minimum in any skill will be taught by teachers who are specially trained in diagnosing the causes of each child's difficulties and in adapting instruction accordingly.

3. It, therefore, has the best chance of bringing

low performers speedily up to the point where they will be able to rejoin pupils in the regular class.

The principal arguments against this approach are:

1. It relies too heavily on a single cut-off score which, because of the inevitable error of measurement, may assign pupils to remedial classes who don't need such instruction, and overlook other students who do.

2. It focuses attention on certain students as "failures" and thereby is apt to damage their self-esteem to the point where they may see themselves forever as "losers."

3. It is often, in effect, a device for segregating minority children within the same school and thereby heightening the tensions between them and their contemporaries in the regular classes.

For deciding which students should repeat the same grade: In this case, the testing for minimum competence is generally an end-of-year effort mandated by the school district or by state law. The aim is to discourage the practice of "social promotion" and the tendency of some teachers to give out passing marks to all their students. The assumption is that if a student turns in test scores that the authorities have decided are below the acceptable minimum for his or her age level, he or she will be helped by going over the same ground for another year.

The principal arguments in favor of using criterion-referenced tests for this purpose are as follows:

1. They ensure that the specific objectives of instruction in each grade will be crystal clear to every teacher in every grade in all the schools.

2. They provide a definite basis for holding every teacher accountable for teaching toward these specifics.

3. They also provide a definite basis for holding every student accountable for mastering the basic skills appropriate to each grade level.

The principal arguments against this practice are as follows:

1. It forces teachers to concentrate on students who are likely to fall short of the minimum and to neglect the needs of those who may be well beyond the minimum.

2. It makes the assumption that the higher authorities in the system have the knowledge to determine what constitutes minimum competence in all skills at each age level.

3. It is so destructive of the self-esteem of children left behind that they tend to become permanently discouraged and fall farther and farther behind as they grow older.

Clearly, these issues are not easily resolved, and they are likely to be a bone of contention in many school districts for a long time to come. Because of the diversity of problems and students in each district, the issues in question are probably more likely to be resolved at the local level than at the state level, provided that more parents are willing to give attention to school policy problems than they have in the past.

For deciding which students shall be denied a high school diploma: This use of minimum competency testing is unquestionably the one that has been gaining the greatest prominence in recent years. It is also the one that has stirred up the greatest political controversy, as one state after another has adopted some form of minimum competency testing as a high school graduation requirement or is contemplating doing so.

Most of the arguments we've just listed for and against the other uses of minimum competency testing also apply to its use as a graduation requirement. There are, however, some additional arguments, both pro and con, that are argued about this point with particular vehemence.

The ones in favor are as follows:

1. By stiffening graduation requirements through the use of minimum competency tests, we shall ensure that fewer students will loaf their way through the schools without mastering the skills they will need to become self-supporting and productive citizens.

2. Schools owe employers and/or colleges the right to know that the award of a high school diploma guarantees that its possessor has at least some of the skills needed to carry on beyond high school.

And the opposition arguments are:

1. No one is yet wise enough or knowledgeable enough to frame a test that encompasses all the skills that may be minimally required for survival in

college or the job market, or in any other aspects of life.

2. The setting of minimum standards in terms of scores on any kind of test is a matter of human judgment, and there is as yet no way of guaranteeing that such judgments will not be so infected with personal bias (wittingly or unwittingly) that they will adversely affect the lifetime chances of many students, particularly those in minority groups.

This controversy is heating up. In fact, the use of minimum competency tests to hold students back or deprive them of a diploma has become a legal issue that courts of law are being asked to decide.

CHAPTER 5

Questions to Ask About Local Testing Matters

No two schools use exactly the same tests in exactly the same way or encounter exactly the same problems in their use. Because of this, the only way to get a reasonably accurate account of the testing that affects your children is to ask local school personnel specific questions about their testing operations.

The way such questions should be handled will vary from school to school. You might form a series of round-table conversations with the teachers, students, school administrators, guidance counselors, and school psychologists who are in touch with the testing carried on in your child's school. Or you might schedule a series of informal private interviews with many of the same people. However you handle it, every effort should be made to prevent the give-and-take from becoming a threatening experience for anyone. Your goal should be to get the essential facts, not to impugn motives, assess blame, or to start arguments for and against any of the various aspects of testing as it goes on in your local school.

The kinds of questions you might ask are suggested below in four groups:

1. Questions about the tests that the teachers prepare for their own students.
2. Questions about the tests that are purchased from publishers for use in a given school or school district.
3. Questions about the tests (if any) that the local school administers on behalf of the state education authority.
4. Questions about miscellaneous local testing matters.

Questions About the Teachers' Own Tests

How many and how time consuming? Not much is known about the tests that teachers prepare for their own students. Yet it is conceivable that, in many schools, this kind of testing is the most prevalent and consumes more student time than all other kinds put together. Therefore, you might start off by asking the teachers to estimate how many tests of their own making they ordinarily give during a school year, and how much classroom time the students spend taking their tests.

What for? In Chapter 1, we looked at ten possible reasons for giving tests in the schools. For most teachers, the first three reasons will probably be the most applicable: (1) to help motivate student learning; (2) to provide feedback for individualizing instruction; and (3) to help make placement decisions about their students—for example, to help decide whether a student should be assigned to a special class for remedial instruction. It might be a good idea to ask the teachers which one of these three reasons they regard as most important and which one as least important. They might then be urged to state any other reasons they have for giving tests.

What types of tests? To get an idea of the types of tests the teachers make up, look at some samples of the tests they have recently prepared. The type of test used will undoubtedly vary a good deal depending on the subject matter, the students' attributes being tested, the grade level at which the teachers are working, and the teachers' own preferences. Many tests are likely to be short and quite informal. Most will probably aim at assessing students' cognitive attributes, but some may attempt to get at affective attributes, such as the students' preferences for TV programs, how they feel about themselves and their classmates, etc. You will also want to know how many of the teachers' tests are of the true-false or multiple-choice variety, and how many are of the free-response or essay variety. The goal here is for you to get an idea of which types of test exercises are given most often.

There is a possibility that, in pursuing this line of questioning, some confusion may arise between what constitutes a *test* exercise and what constitutes a *learning* exercise. One way around

this difficulty would be to define a test exercise as any student activity *that the teacher prescribes for the express purpose of making and recording a judgment (in the form of a grade or score) about the quality of the student's performance.*

How are the tests prepared? Some teacher-made tests will be "quickies," possibly given orally and off-the-cuff. Others will be more formally designed to cover extensive units of work and may be adapted from material in the students' textbooks or workbooks. You might ask how much the teachers in the school help one another in planning and reviewing the content of these more formal tests before presenting them to their students. You might also ask how much help they get in such test preparation from curriculum supervisors or from the district director of testing.

The basic question to be asked about test preparation is what the teachers do to ensure that their classroom tests are *content valid*. That is, how do they determine whether and to what extent the test exercises they make up, or borrow from other sources, are in line with the agreed-upon objectives of the subjects they are teaching, or of the attitudes they are hoping to develop in their students?

It might also be helpful to ask two incidental questions: (1) Has the teacher ever taken a course or attended a workshop on test construction? (2) If so, has he or she found such a course or workshop of any practical help in devising tests?

How are the tests graded or otherwise rated? The grading of student work has been the subject of heated argument lately, not only about how it should be done, but whether it should be done at all. Most teachers will probably agree that some form of grading is necessary as a way of periodically informing you and your child about how well your child is doing and about what particular aspects of his or her schoolwork need improvement.

The crucial question here is whether the teacher tends to grade tests in terms of *degree of mastery* or in terms of *relative performance*. In other words, is the test grade supposed to indicate how close your child has come to what the teacher regards as perfect mastery of whatever skill is being tested? Or is it supposed to indicate how your child's test performance ranks with that of the other students in the class? Or does the teacher use some combination of these two approaches?

Another crucial question has to do with the *reliability* of the test grades or scores. Do the teachers invite their colleagues to check their grading of a batch of tests from time to time to see how closely they agree in judging the quality of their students' performance?

Questions About Tests Purchased From Test Publishers

How often given and how time consuming? Perhaps the best way to approach this question is to ask teachers whether, in the course of a school year, this kind of testing takes up more or less student time than the tests they make up themselves. You might then ask: (1) How many times during the school year do the students in a given grade take such tests? and (2) How long on the average does each testing take?

What for? The members of a school staff probably have a number of reasons for buying and giving commercially available tests. To begin to get these reasons out in the open, you might question the school administrator about what purposes he or she thinks such tests serve; also, what purposes they do *not* serve. Similar questions might be asked of the teachers, the guidance counselor or school psychologist, and the students.

What types of tests? A good way to begin discussion of the types of tests a school or school district buys might be to refer the staff to the four categories of tests we discussed in Chapter 1, and to ask which type is most heavily emphasized in the formal testing program and which one is least emphasized.

You might then look into the kinds of questions each type of test contains. How much reliance is placed on objective type questions (true-false, multiple-choice, etc.) and how much on free-response type questions (sentence completion, essay questions, etc.)?

As with teacher-made tests, the best way to get a quick impression of the testing tasks your children are given is to examine a sample of the test questions and try answering them yourself. This experience should lead to a host of specific questions about the kinds of pupil learning each type of question is supposed to be getting at, and how well the test as a whole may be measuring such learning.

How and by whom were the tests selected? Any school or school system that tries to purchase educational tests best suited to its own educational purposes faces a massive problem because of the vast number of tests for sale. One way to open up the questioning on this matter might be to present the school staff with three multiple-choice questions, somewhat like these:

1. Who was *primarily* responsible for deciding which commercially-prepared tests were most recently bought for use in this school?

- (a) The school principal
- (b) The school psychologist
- (c) The supervisor of curriculum
- (d) A committee of classroom teachers
- (e) Other

2. What was the *first* step in narrowing down the number of different publishers' tests to be considered in detail?

- (a) Talking with the publishers' salespersons
- (b) Reviewing publishers' test catalogs
- (c) Consulting the test reviews in professional periodicals and/or in the *Mental Measurements Yearbooks* (Buros 1979)
- (d) Consulting colleagues in other school districts
- (e) Other

3. What was the *chief* consideration in deciding on which one among several publishers' offerings to buy?

- (a) The closeness of fit between the tests' content and the school's curricular objectives
- (b) The quality of service the test publisher is able to provide
- (c) The ease with which the tests can be administered
- (d) The price
- (e) Other

How are the tests administered? The manner in which a test is administered might affect the validity of the test results. Most published tests, therefore, are accompanied by manuals giving elaborate directions for how they should be given. These directions try to ensure that all schools follow the same test-giving procedures as closely as possible. However, the degree to which the prescribed procedures are followed in a particular

school may vary considerably depending on who actually administers the tests and on the setup in the rooms where they are given.

Thus, your line of questioning on this point could go something like this:

1. Where are the tests usually administered? In the classrooms? In the auditorium? In the cafeteria?

2. Who administers the tests? Specially trained supervisors? Supervisors assisted by classroom teachers? Classroom teachers only? Others?

3. To what extent are the procedures rehearsed in advance of the actual administration?

4. What measures are taken to ensure that disturbances in the testing room will be kept to a minimum?

5. What measures are taken to prevent and/or discourage cheating?

6. What is done in advance of the testing to ensure that the students will be familiar with the types of questions the tests contain?

7. How did the students themselves feel about the test administration? Were they clear about how to proceed? Or were students confused? Did they feel they were able to do their best? Were they bored or bothered by the whole proceeding?

How are the tests scored and reported? Most of the large test publishers these days provide a service whereby their tests are scored by machine and reported back to each school on classroom rosters. A school usually has the option whether or not to purchase this service. Questions on this and related matters could go as follows:

1. Does the school ordinarily purchase the machine-scoring service? If so, why? If not, why not?

2. If the publisher's scoring service is used, how much time elapses between the giving of the tests and receipt of the score reports?

3. If the tests are scored by hand locally, who does the scoring? What provisions are made to ensure its accuracy?

4. Do the score reports provide detailed information that show which particular questions each student got right or wrong? If so, do teachers find this information useful? How useful? In what ways is it used?

5. What method is used for converting raw

scores to scaled scores? Percent correct? Percentile rank? Standard scores? Grade equivalents? Mental age? IQ scores? Other?

How are the test scores interpreted and used?

This question is, of course, the most important one to be asked of any school staff, since there is no point in giving commercially available tests at all if the scores are simply filed away unread and unused.

You might open up a discussion of this matter of interpretation and use by asking questions like the following:

1. In what form are the score reports transmitted to students, teachers, parents?

2. Who in the school is responsible for making sure that the people who see the scores understand what they mean (a) for each child? (b) for each teacher's class as a whole? (c) for the school as a whole?

3. What is done to ensure that any educational decisions based on the scores will take due account of the sampling error (the standard error), which is embedded in every test score?

4. Are the scores used to help each student evaluate his or her progress and understand how he or she might do better? Make plans for his or her future education? If so, how is this done? By whom?

5. Are the scores used to help teachers evaluate their own classroom procedures? Their own grading of individual students? If so, how is this done?

6. Are the scores used in any way to help the school staff evaluate the educational effectiveness of the school as a whole? If so, how is this done?

Questions About Tests Supplied by the State Education Authority

The first question you should ask about tests in this category is: Does the state education authority have any kind of testing program aimed at the schools within its jurisdiction? If the answer is "Yes," your line of questioning might then go somewhat as follows:

1. What is the nature of the state's testing program?

2. What subjects in what grades does it cover?
3. What types of tests does the state supply?
Norm-referenced? Criterion-referenced? Objective*? Free-response? Other?

4. How are the tests intended to be used? To test for: (a) minimum competence? (b) promotion from grade to grade? (c) eligibility for high school diploma? (d) "accountability" of local school personnel? (e) allocating state funds to local school districts? (f) the guidance of students? (g) admission of students to state colleges and universities?

5. To what extent *are* the tests actually used for any of the above purposes?

6. Are some students exempted from the program? If so, under what conditions?

7. Is the state testing program mandatory for all schools in all school districts? Or is it optional in whole or in part?

8. If there are any options, which ones do your school or school district choose to exercise?

9. At what time of the year are the tests given? How much time do they consume?

10. Who actually administers the tests? (a) state supervisors? (b) local supervisors? (c) teachers? (d) others?

11. How are the tests scored? By whom?

12. In what form are the test results reported to the local school?

13. To what extent are the test results of individuals made available (a) to the students themselves? (b) to their teachers? (c) to their parents? (d) to the local education authority? (e) to the news media?

Another line of questioning on the state testing program (if there is one) might take the following tack:

1. To what extent do local school personnel participate in the development and/or the selection of the tests?

2. Who finally decides on what specific questions will be included in each of the tests?

3. What evidence is available on the validity of the tests?

4. What evidence is available on the reliability of the tests?

5. What evidence is available on the degree to which the tests may be biased against certain groups of students?

Questions About Miscellaneous Local Testing Matters

The foregoing three sets of questions are intended to encompass the kinds of educational testing that are likely to have the most impact on your children. There are, however, other kinds of testing that may figure more prominently in some schools than in others. Examples are given below, together with a few questions that you can ask about each.

Individually administered IQ tests

1. Are such tests administered to students in your school? If so,
 2. Are the persons who administer and interpret the tests properly trained to do so?
 3. At what ages are the pupils tested?
 4. How are the results of the tests used?
 5. Who has access to the results of the tests?
 6. What evidence is locally available on the validity of the tests? On their reliability? On the degree to which they may be biased against certain groups of children?
 7. If there has been evidence of bias, what is being done to correct the situation?
 8. What use is made of the IQ scores? By teachers? By others in the school system?
 9. Are the IQ scores entered on the students' cumulative records?
 10. If so, are they removed from the record after a certain period of time? How long a time?

Tests of attitudes, interests, self-concept, etc.

1. To what extent are such tests given to the pupils in your school?
 2. How are they administered? In groups or individually?
 3. How are the tests used? For individual counseling? For group counseling? Other?
 4. What specific tests are used, and why were these particular tests chosen?
 5. What evidence is locally available on the validity of the tests? On their reliability?
 6. Who has access to the results of such tests?
 7. Are the results entered on the students' cumulative records?
 8. Are some students exempted from taking such tests? If so, why?

College admission tests

1. How many students in your local high school normally go on to college? What percent of the graduating class does so?

2. How many of those who are headed for college ordinarily take college admission tests of one kind or another?

3. For what college admission testing programs do they register? Those of the College Board? The American College Testing Program? Other?

4. What is done by local school personnel to help students in the process of registering for the tests? In qualifying for waivers of test fees when appropriate?

5. What, if anything, is done to help students prepare for the tests? Practicing on similar tests provided by the publishers or others? Setting up special classes for the purpose? What evidence is there that such special classes have been helpful to the college candidates?

6. Are any of the college entrance testing programs administered in the local high school? By whom?

7. To what extent does the school use the results of college entrance tests for assessing its own educational operations? How? And why? Or why not?

8. How do the results of such tests compare (a) with the grades the teachers have given the same students? (b) with the results of the school's own testing program?

9. Is the school calendar set up to avoid conflicts with major testing dates?

The questions in this chapter do not, of course, exhaust all the questions that *could* be asked about testing in your local schools. They are merely intended to suggest the *kinds* of questions you might ask in order to get a reasonably complete picture of the impact of local testing operations on your children. We hope that, as the picture emerges, you can work with school personnel to find ways of ensuring that the testing that goes on in your schools will be handled so that it enhances the quality of your children's education.

Appendix A

State Activity in Minimum Competency Testing

The table that follows was prepared by Chris Pipho, Associate Director, Department of Research and Information, Education Commission of the States, Denver, Colorado. It is reprinted here with Mr. Pipho's permission.

The table summarizes state activity during the period 1972-1978. The table was published by the Commission on October 5, 1978. Information on thirty-six of the fifty states is provided. The fact that data is missing on the remaining fourteen states does not necessarily indicate that there is no current or contemplated program in those states.

For an earlier survey of state testing activity in all fifty states, see *Basic Skills Assessment Around the Nation*, Princeton, N.J.: Educational Testing Service, September 1977, 24 pp.

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
ALABAMA State Board Resolu- tion 1977	*State Department of Education	9th-12th	*Basic skills state test	For high-school graduation	*The State Superintendent of Education has appointed a committee of 100 to study and make recommendations on standards, the test and high-school diploma issue. These recommendations are due 7-1-78.
ARIZONA State Dept. of Edu- cation ruling	Set by the State Dept. of Educ. assessment by local district; no statewide test prescribed	8th grade 12th grade	Reading, writing, computation	For graduation from 8th and 12th grades	Effective January 1, 1976, students must be able to read, write, and compute at the 6th-grade level to receive the standard 8th- grade certificate and at the 9th-grade level for high-school graduation.
CALIFORNIA (early exit test) SB 1112-1972 SB 1243-1975 SB 1502-1976	Test and cut-off standards set by the California Depart- ment of Education	Age 16 to adult	Process areas of reading and compu- tation content areas: consumer econom- ics and mathematics	Award high-school diploma and leave school with parent permission	Title of program: California High-School Proficiency Test. Test administration and security handled under contract with private testing corpora- tion.

AB 3408-Ch. 856-1976	State Board to supply performance indicators and examples of minimum standards. Local districts set graduation standards	Test once between 7-9 and twice between grades 10-11	Reading, writing, computation	For high-school graduation	The act also prescribes course of study requirements. In effect for the graduation class of 1980.
AB 65, Ch. 894	Local districts	Once between grades 4-6, once between grades 7-9; twice between grades 10 and 11	Reading comprehension, writing, computation skills	For remediation and grade promotion decisions	This law is to move California into elementary grade testing.
COLORADO SB 180-1975 C.R.S. 22-32-109.5	Local boards	9-12—test twice a year	Local district option	Local district option for high-school graduation	If a local board of education imposes a proficiency test for high-school graduation, it must follow the stipulations of this act.
CONNECTICUT State Board Resolution, Nov. 1977 (Early exit program)	State Department using existing GED and APL test standards	16-year-olds	Academic: General Educational Development (GED) test. Life Skills: Adult Performance Level (APL) test	With appropriate counseling and follow-up teaching, students can leave school with a high-school equivalency diploma	Final implementation contingent upon \$300,000 appropriation from legislature in 1978.

Appendix B

Glossary of Terms

The definitions contained in this glossary are not intended as general definitions of the terms used, but as definitions specific to the educational testing field. A few terms are included that do not occur in the main text of this handbook, but that may come up in discussions with your local school personnel.

Achievement test: Any prescribed sampling of student behavior, the purpose of which is to make an estimate of what students have learned in a given subject up to the time the sample is taken.

Affective attribute: That characteristic of student behavior pertaining to a student's feelings or emotions.

Aptitude test: Any prescribed sampling of a student's behavior, the purpose of which is to predict how well a student may function later on in some activity.

Attitude scale: A device for estimating the degree to which a student has a favorable or unfavorable feeling toward something or somebody.

Example: How important is it to you to do your homework well?

- A. I think doing good homework is not at all important.
- B. I think doing good homework is important but not as important as other things.
- C. I think doing good homework has about the same importance as other things.
- D. I think doing good homework is more important than most things.
- E. I think doing good homework is more important than anything else.

Average (or mean): The average, or mean, of a group of numbers is the sum of the numbers divided by the number of numbers in the group.

Example: The average of the three numbers 1, 6, and 5 is 4, since the sum of $1 + 6 + 5$ is 12, and 12 divided by 3 equals 4.

Basic skill: Any skill that is thought to be prereq-

visit either (1) to further learning or (2) to coping with the demands of life.

Battery of tests: A group of tests—usually standardized achievement tests—each of which samples the student's learning in one particular school subject. The results from such a battery are often presented in the form of a profile, showing how the student's performance in each area of subject matter compares with his or her performance in the other subject matters sampled by the tests in the battery.

Bias in testing: A test is biased if there is a systematic error in the results that leads to a mistaken inference concerning the present or probable future behavior of a student or a group of students. (See "random error" and "systematic error.")

Cognitive attribute: That characteristic of student behavior pertaining to how much a student knows and/or how well the student knows how to do something.

Construct validity: That characteristic of a test that enables you to infer from the scores the degree to which the test takers differ with respect to some psychologically defined quality, such as creativity, self-esteem, mathematical intuition, and the like.

Content validity: The degree to which the kinds of tasks sampled by a test conform to the objectives of a particular instructional program.

Correlation coefficient (r): An index of how closely two different measures (such as test scores and school grades) agree in the way they rank the performance of students.

Criterion-referenced test: A test that is scored in such a way as to indicate how closely a student's performance meets specifications as defined by the nature of the tasks contained in the test itself. It is usually scored in terms of the number or the percentage of the tasks the student succeeds in performing properly. Some fixed percentage (e.g., 80 percent, 100 percent) is taken to mean that the student has mastered tasks of the type tested—i.e., that the student has met the "criterion" set by the test maker.

Example: In carpentry, students might be given a criterion-referenced test in the use of a hammer. The test might consist of driving twenty nails. The criterion score (passing score) on the

test might be set at 16. That is, if the student had driven sixteen of the nails without damaging the board or bending any of the nails, he or she might be considered to have "passed" the nail driving test. The pass criterion might be set higher or lower depending on how certain the test giver wants to be that the student can accurately drive home any specified number of nails.

Cumulative record: A record that contains the scores on all the tests a student has taken from the time he or she has entered school.

Deviation IQ: The type of IQ score that is expressed as a deviation from the mean of some standard reference group.

Example: The mean of the standard reference group for IQ scoring is set at 100. If a student's IQ score is 105, this is taken to mean that the student's score deviates from the mean of all students of the same age by 5 points. If another student's IQ score is 95, this is taken to mean that this student's score deviates from the mean in the opposite direction by 5 points, and the difference between the two students' scores would thus be 10 points. (Compare "Ratio IQ").

Educational test: A transaction between a test giver and a student, which is designed to yield a sample of the student's responses to a sample of tasks that are prescribed by a test maker and that, in the aggregate, are intended to form the basis for rating the student's behavior with respect to some attribute.

Equivalency test: A test designed for the granting of academic credit to students who have had no formal high-school or college instruction in the subject tested.

Essay test: A free-response test in which students are expected to organize their answers in sentences and paragraphs of their own devising.

Examiner reliability: The degree to which two or more examiners agree in their ratings of students' responses to a test situation. (This is sometimes referred to as "inter-observer reliability.") It also includes the degree to which an examiner's ratings on one occasion agree with his or her ratings of the same set of responses on another occasion. (This latter is often referred to as "intra-rater reliability.")

Free-response test: A test containing tasks to

ACTION TAKEN	SETTING OF STANDARD ^a	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
DELAWARE State Board of Education Resolu- tion 1976	State Department of Education, after baseline test is given	Baseline instrument grade 11 1978	Level II Competen- cies, i.e.: the appli- cation reading, writ- ing, and mathe- matics	For high-school graduation	The development of the baseline testing instrument has been awarded to a private test contractor.
FLORIDA CSSB 340 1975 F.S. 229.814 CSSB 107 1976 F.S. 76.226	State and local boards	Grades 3, 5, 8, and 11	Basic skill areas Functional literacy	For high-school graduation, grade promotion, and early exit from high school	Titled: The Educational Accountability Act. Places ban on social promotion.
GEORGIA State Board of Education action	State Board and State Department	Grades 4, 8, and 11	Grades 4 and 8: read- ing, math, social studies, and science. H.S. test under development	The State Board of Education has not taken final action on the use of the test's	The student assessment program is part of a long-range plan to fund local school dis- tricts through a program of comprehensive planning.
IDAHO State Board of Education resolution	State Board of Education	Grades 9-12	Reading, writing, arithmetic, and spelling	For high-school graduation at the option of local districts	Students passing the proficiency program will receive a diploma with the state board seal. In districts not participating, students will receive district-diploma.

ILLINOIS
HB 238

Local districts with assistance from the State Board of Education

Local districts with assistance from the State Board of Education

Local districts with assistance from the State Board of Education

Local districts with assistance from the State Board of Education

This act provides for the State Board to prepare procedures and materials to encourage and assist local school districts to develop minimal competency testing programs by December 15, 1978. The State Board of Education is to report back to the Legislature on state progress by June 30, 1980.

INDIANA

State Board Resolution adopted February 1978

To be set by local districts with advisory committee of teachers, administrators, parents, and community members

3, 6, and 10

Reading

Primarily for remediation; local district may use for other purposes

Rules and regulations supporting this resolution awaiting Attorney General's and Governor's approval (3/14/78).

KANSAS

State Board of Education adopted policy statement, January 1978

State Dept. of Education will set standards and goals with assistance of local districts

3, 6, 9, and 12

Academic skills: reading and mathematics, grades 3, 6, and 9.
Life skills: grade 12

Local district option

Legislation is pending that has a good chance of passing yet in 1978. This policy statement did not include a program of implementation. Enacted law will supercede.

NCCE/61

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
KANSAS (cont.) Senate Substitute for HB 3115 (passed in April 1978)	State Board of Edu- cation	Grades 2, 4, 6, 8, and 11	Reading and mathe- matics	Not mentioned in law	Law calls for a voluntary participation in '78-'79 school year, full district participa- tion in '79-'80. At the close of the '79-'80 school year, results of the pilot effort are to be reported back to the Governor and the Legislature.
KENTUCKY State Board of Edu- cation adopted 4-yr. competency plan in 1977	State Department with task force assistance	Grades 3, 5, 8, and 11	Criterion-referenced test in reading, writing, and arith- metic	For high-school graduation and grade promotion	
HB 579, 1978	Local districts are to develop educational improvement plans	Grades 3, 5, 7, and 10	Reading, writing, spelling, language arts, and mathe- matics	Not mentioned in law; presumed to be a part of local plans	HB 579, the Education Improvement Act, also calls for the appointment of a state- wide task force and for local districts to include student remediation in their plan. Beginning 1979.
LOUISIANA HB 810, Act 709, 1976	State Superintendent of Schools	Grades 4, 8, and 11	Basic communica- tion and computa- tional skill areas with criterion-refer- enced test	No mention of grad- uation or grade pro- motion	Requirements a part of the accountability and assessment law.

MAINE

Enacted amended
LD 1810, Ch. 78 of
the Private Special
Laws

State Department of
Education

Grade

(See comments)

For high-school
graduation

The law specifies that the Commissioner of Education is to be assisted in the development of the assessment instrument and the areas to be tested by a statewide committee. The law names the organizations to be represented.

MARYLAND

HB 1433, M.A.C.,
Art. 77, Sec. 980,
1976 and HB 1462,
Ch. 559 1977

State Board of
Education

Grades 3, 7, 9, and
11

Reading

For grade promotion
and high-school
graduation

Students not meeting minimum requirements may be retained in same grade or enrolled in remedial program.

MASSACHUSETTS

State Board of Education

Local school districts; minimum standards for basic skill competency by September 1980

Early elementary,
later elementary,
secondary

Mathematics and
communication
skills

"The purpose is *NOT*
to establish a new
condition for promotion
or graduation."

Intent of the board policy is to . . . "improve basic skills competency in partnership with local school officials."

Test: Elementary, local option Secondary options:

- (1) State test
- (2) Commercial test approved by State Department of Education
- (3) Local test approved by State Dept. of Education

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
MICHIGAN State Department of Education	State Department of Education	Grades 4, 7, and 10	Reading and math	For local district use	12th-grade minimal competency test, covering life role skills under study.
MISSOURI State Board of Education	State Department of Education	Grade 8	The application of reading, mathematics, and government/ economic skills	No mandate for graduation or grade promotion	The Basic Education Skills test, developed by the Department of Education, has been field tested in grades 8, 10, and 12.
NEBRASKA State Department of Education	Local school districts	Begins in grade 5 and continues until mastery is achieved by each student in each skill area	Reading, writing, and mathematics	State developed test <i>is not</i> to be used for grade retention or promotion	Local school districts may use the Nebraska Assessment Battery of Essential Learning Skills (N-ABELS) or develop their own test.
NEVADA AB 400-1977	State Board of Education	Grades 6, 9, and 12	Reading, writing, and mathematics	For high-school graduation	Students may be promoted to the next grade if promotion is failed but remedial work is to be provided.
NEW HAMPSHIRE State Department of Education	State Department of Education	Grade 4 Grade 8 and 12 under development	Communications and mathematics	Grade promotion and high-school graduation <i>not mentioned</i>	Program is to serve as guidelines for local school districts.

NEW JERSEY

A. 1736, Ch. 97 1976	Statewide committees	—	Reading and mathematics	Student diagnosis and remedial identification	Local districts are to provide remedial assistance in order for students to meet state standards.
State Department of Education	State Department of Education	3, 6, 9, and 11	Reading and mathematics, Life skills under study	Identification of students needing instructional assistance	Development of testing instrument is under contract to a private test development company.

NEW MEXICO

State Board of Education	State Department of Education	Elementary level; grade choice up to local district and grade 10	Elementary: Local option Secondary: Proficiency battery based on adult performance level (APL) and a writing sample	Proficiency endorsement on high-school diploma if test is passed	The state department of education is working on a series of basic skill area curriculum guides.
--------------------------	-------------------------------	--	--	--	---

NEW YORK

State Board of Regents amended Section 103.2 of the Commissioner's regulations, August 1978	State Board of Regents	Grades 11 and 12	Regents' competency test: reading comprehension, writing, and mathematics	For high-school graduation	The changes made by the Board of Regents are designed to create a comprehensive program for early identification of students needing remedial assistance and to correlate the requirements for competency in reading, writing, and mathematics between the high school Regents' examination and the Regents' competency testing program.
		Grades K-12	Formal and informal assessment techniques	Early identification of students needing assistance	

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
NEW YORK (cont.)		Grades 3 and 6	Pupil evaluation program: tests in reading and mathematics	Early identification of students needing assistance	
		Grade 8 or 9	Preliminary competency tests: reading and writing	Early identification of students needing assistance	
NORTH CAROLINA Enacted HB 204	By Competency Test Commission created by law	Grade 11 beginning in 1978-79	To be determined by the Competency Test Commission; see Comments	For high-school graduation	The test is to measure "those skills and that knowledge thought necessary to enable an individual to function independently and successfully in assuming the responsibilities of citizenship."
Enacted HB 205	By Competency Test Commission (separate from HB 204)	Grades 1 and 2 criterion-referenced test Grades 3, 6, and 9 norm-referenced test	To be determined by the Competency Test Commission	To be determined by the Competency Test Commission	

OREGON

State Board of Education 1972

Local Districts

District Option

Reading, writing, and computation. Local option for personal development, social responsibility, and career development

For high-school graduation

Graduation requirements are based on course credit, attendance and required competencies in personal, social, and career areas.

OKLAHOMA

State Board of Education

Decision will follow baseline data test

Grades 3, 6, 9

Reading

Decision will follow baseline data test

The baseline test instrument was administered to 15,000 students in the fall of 1977.

Grade 12

Survival skills

RHODE ISLAND

State Department of Education

Grades 4, 8, and age 17 used in pilot test

Comprehension analysis and evaluation known as the "life skills"

For local district and state department use. Competency-based, high-school diploma under study

The Rhode Island master plan for evaluation and the statewide assessment program have jointly contributed to the pilot testing in the life skills areas.

Board of Regents

Board of Regents in conjunction with State Standards Council

Elementary, early and late secondary grades

Elementary: basic skills.
Secondary: Minimum competency standards in selected life skills

—

This policy was adopted 5/11/78 to support the master plan for evaluation mentioned above. The policy also calls for the development of standards of excellence to be "measured by extensive assessment of advanced life skills, scholastic, and cultural achievement for students aspiring to attain recognition by the regents for special distinction along a variety of dimensions."

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
SOUTH CAROLINA Enacted H 2053 August 1978	State Board of Education	1, 2, 3, 6 and 8: Basic skill areas. Grade 11: Adult functional competency	Mathematics and communication, skills of reading and writing	For remediation (See Comments)	—Also establishes Basic Skills Advisory Commission, 15 members. —Provides for Grade 1 readiness test for students entering school. —Grade 11 high-school test will be used to gather baseline data until 1989. At that time, the state board is to make a decision regarding high-school graduation and the test.
TENNESSEE State Board of Education Nov. 10, 1977	State Department of Education	Grades 11 and 12	Reading, math, grammar, and spelling	For high-school graduation, class of 1982	Two different programs were established in the same court ruling.
State Board of Education Nov. 10, 1977	Local districts	Grades 4, 5, 6, and 8	Subjects at discretion of local district	Remediation	
UTAH State Board of Education January 1977	Local district	Local district	Reading, writing, speaking, listening, arithmetic, democratic governance, consumerism, problem solving	For high-school graduation, class of 1980	Local districts will choose the test and set the cut-off standards.

VERMONT

State Board of Education

State Board of Education

Continuous

Reading, writing, speaking, listening, and mathematics. Area of reasoning is under study

For high-school graduation

Mastery of the basic competencies is to be one requirement for graduation unless student has a limiting handicap.

VIRGINIA

HB 402, 1978

State—reading and mathematics

Grades 9-12

State—reading and mathematics

For high-school graduation beginning with class of 1981 and for planning for specific instructional needs

Testing in reading and mathematics of 9th and 10th grades in school year of 1978-79 for initial diagnosis for instructional needs.

Local—essential citizenship skills*Local*—essential citizenship skills

Subsequent years, reading and mathematics testing to begin in 9th grade.

Skills to qualify for further education or employment

Skills to qualify for further education or employment

Local assessments of essential citizenship skills and skills to qualify for further education or employment can be by means other than tests.

Initiated in HB 256, 1976

State Board of Education in cooperation with local districts

Grades K-6

Reading, communications, and mathematics

To measure progress of individual students

Field testing in spring of 1978. Testing for all students K-6 to begin in school year of 1978-79.

Repeated in HB 402, 1978

ACTION TAKEN	SETTING OF STANDARDS	GRADE LEVELS ASSESSED	SKILL AREAS ASSESSED	USE OF STANDARDS AND TEST	COMMENTS
WASHINGTON SB 3026 HB 1345-1976	Local districts	4th and 8th	Standardized achievement test in reading, mathematics, and language arts	For school district and parent use	School districts are encouraged to develop a separate test for second grade for the early identification of pupils needing assistance in language and computational skills.
WYOMING State Board of Education policy	Local districts will establish specific standards	—	Reading, writing, and computing with proficiency. Process and structure of democratic governance and free enterprise	For high-school graduation	The start-up date and rules and regulations are to be drawn up by the State Department of Education. Local districts will verify the state standards and are encouraged to go beyond the minimum standards set by the State Board of Education.

which the student responds by supplying words or other actions that are not supplied in the test itself.

Examples:

1. What is the northernmost state of the United States?

2. Here is a quarter. Please give me the same value by picking out from this set of coins the right number of pennies, nickels, and dimes.

3. What is meant by the expression, "More honored in the breach than in the observance?" Illustrate by writing a story giving an example of how the expression might be used.

Frequency distribution (of test results): A table that shows the number of students obtaining each score or grade on a given test.

Example: Distribution of scores on a fifteen-item spelling test taken by fifty, fourth-grade students.

Score (Number of words spelled right)	Number of students obtaining each score
15	2
14	0
13	3
12	3
11	5
10	6
9	6
8	7
7	6
6	4
5	2
4	1
3	2
2	2
1	0
0	1
Total = 50	

Grade equivalent score: A scaled score derived from the raw scores obtained on a test taken by students on each of several grade levels (see "Scaled score"). The grade equivalent score for a particular test is the school grade for which that test performance is the average.

Interest inventory: A questionnaire, the results of which are supposed to show how closely a stu-

dent's preferences match those of persons in each of several educational or occupational fields.

IQ score: A standard score derived from the raw scores on an intelligence test. (See "Deviation IQ" and "Ratio IQ.")

Mastery test: A test designed to show whether a student has successfully performed all the tasks specified in a given learning objective. (See "Basic skill," "Criterion-referenced test," "Minimum competency test.")

Mean score: The mean of a group of scores is the sum of the scores divided by the number of scores in the group. (See "Average.") (Compare "Median.")

Median (50th percentile): The median, or 50th percentile, of a frequency distribution of scores is that score which separates the top half of the scores from the bottom half.

Example: See the frequency of spelling scores after the glossary listing for "Frequency distribution." Here the median spelling score would be 8.5, since this would be the score that would separate the top twenty-five scores from the bottom twenty-five scores.

Mental age score: A scaled score derived from the raw scores made on an intelligence test by students at each year and month of age. It is expressed in age units and corresponds to the test performance that is average for the particular age. This type of scaled score is no longer in general use.

Minimum competency test: A test designed to identify students who are not yet ready (1) for further learning in a given subject or (2) for coping with the demands of life. (See "Basic skill," "Criterion-referenced test," "Mastery test.")

Multiple-choice item: A test item in which the test taker's task is to choose the correct or best answer from several given answers or options.

Examples:

1. After the _____ the sun came out.
 - (a) pretty
 - (b) rain
 - (c) apple
 - (d) ran
 - (e) soon

2. If $\frac{y}{2} = 5$, what is the value of y ?

- (a) $2\frac{1}{2}$
- (b) 5
- (c) 7
- (d) 10
- (e) none of the above

Note: When a multiple-choice item is used to sample a student's feelings, opinions, or beliefs, it is usually presented to the student as having no "right" or "best" answer. *Example:* Which one of the following desserts do you like best?

- (a) cheese cake
- (b) vanilla ice cream
- (c) apple pie
- (d) stewed figs
- (e) animal crackers

Norm: See "Test norms."

Normal frequency distribution (sometimes called the "Bell-shaped curve"): A theoretical frequency distribution of scores or grades that has the following characteristics: (1) the mean score is the same as the median score; (2) the majority of the scores cluster close to the mean; (3) the number of scores beyond any distance above the mean is the same as the number of scores beyond an equivalent distance below the mean; (4) the number of scores becomes increasingly small as the difference between any score and the mean score increases.

Note: An approximation of the theoretically normal distribution of test scores in a given subject produced by any given group of students is usually accomplished by including items that, on the basis of pre-testing, have been found to have the right level of difficulty for the group in question. (See "Test item difficulty.")

Norm-referenced test: A test designed to indicate how a student's behavior compares with that of other students having the same or similar characteristics. Each score on such a test is usually expressed in terms of a percentile rank or in terms of the extent to which it differs from the average (i.e., the norm) of the group in question.

Objective test: A test made up of items for which correct, or best, responses have been determined in advance, so that the raw scores are unaffected by the opinions or judgments of the persons who score the test.

Percent agreement: A rough measure of the degree to which (1) examiners agree in their judgments of an examinee's test performance, or (2) the scores on two or more objective tests rank students in the same order.

Percent correct score: The percentage of test items that a student answers correctly.

Percentile: In a frequency distribution of raw scores, the point below which a certain percentage of the scores fall.

Percentile rank: A number that indicates the percentage of student scores that are at or below a given score in a frequency distribution of scores.

Performance test: A test to which the student is expected to respond by demonstrating how well he or she carries out some prescribed physical act—e.g., playing a piece on the piano.

Personality test: A test, usually in the form of a questionnaire, that purports to sample a student's feelings and beliefs, or behavior tendencies.

Predictive validity: The accuracy with which the scores of a test predict students' future academic or other achievement.

Prescriptive test: A criterion-referenced test, the scores of which are accompanied by specific recommendations about what the student needs to work on next to improve his or her skills.

Random error: The type of error in any test score that is unexplainable except as the result of unknown or chance factors.

Random sample: A sample in which each component among all possible components has had the same chance of being selected for inclusion.

Ratio IQ: An IQ score, now outmoded, that is computed by dividing a student's mental age (MA) by his or her chronological age (CA) and multiplying the result by 100.

Raw score: The score obtained (1) by counting the number of items on a test to which a student has given a correct response or, (2) by counting the correct responses and subtracting some fraction of the incorrect responses so as to discount the effects of blind guessing.

Reliability: An estimate of the degree to which the scores derived from one sampling of student's performance in a given subject would agree with other samplings of the same student's performance in the same subject.

Reliability coefficient: A correlation coefficient that indicates how closely the scores on one test

agree with the scores on another test that samples the same kinds of tasks.

Representative sample: A sample of a total population of students, selected so that each significant subgroup of the population will be proportionately the same in the sample as in the whole population. *Note:* This sampling method rests on the assumption that the proportionate distribution of subgroups in the population as a whole is known in advance.

Sampling error: The error due to random sampling.

Scaled scores: Any set of scores derived from a test in such a way as to make possible rational comparisons with scores from other tests. (See "Standard scores.")

Standard deviation (SD): A statistical index of the amount of spread in the scores the students make on a given test. It is computed by squaring the difference between each score and the mean of all the scores, summing the squared differences, dividing the sum by the number of scores, and taking the square root of the resulting quotient. The formula for the standard deviation is:

$$SD = \sqrt{\frac{\sum (X - M)^2}{N}}$$

where SD stands for the standard deviation; X stands for any score in the frequency distribution; M stands for the mean score of the distribution, and N stands for the number of scores in the distribution. The symbol Σ means "sum of."

If you want to try your hand at calculating a standard deviation according to this formula, you might use the numbers in the frequency distribution of raw scores on the fifteen-item spelling test shown above under "Frequency distribution." For that distribution, you should find that the mean raw score comes to 8.4, and the standard deviation comes to 3.26.

Standard error of measurement (SEM): An estimate of the degree to which an individual student's score might have varied if he or she could have been tested on a series of random samples of the same kinds of tasks. You can get an approximation of the SEM in terms of raw scores by using the following formula:

$$SEM = .432 \times \sqrt{n_i}$$

where n_i stands for the numbers of items in the test. Thus, you should find that the SEM for that fifteen-item spelling test described under "Frequency distribution" comes to about 1.67 raw score points.

Interpretation of the SEM: Suppose a student gets a raw score of 10 on that same 15-item spelling test which has an estimated SEM of 1.67. This would suggest that the odds are a little better than two-to-one that his or her "true score" on such a test is probably no lower than 8 and no higher than 12. (See "true score" below.)

Standard reference group: A group of students that has certain specified characteristics and which has been tested for the purpose of obtaining the data needed for converting raw scores into scaled scores.

Standard score: A score expressed in terms of the standard deviation of scores generated by giving the test to students in a standard reference group.

Example: If the mean score of the standard reference group is set at 50 and the standard deviation of the scores of that group is set at 10, then a student getting a score one-half of a standard deviation above the mean of the standard reference group would have a standard score of 55.

Standardized test: Any test that (1) has been given under specified conditions to the members of a standard reference group having specified characteristics, and that (2) provides a scaled score of some kind. Most standardized tests are accompanied by norms of some kind, in which case they are also norm-referenced tests. The normative data, however, may or may not be derived from testing the standard reference group.

Systematic error: The type of error in a test score that makes the student's score higher or lower than one would expect from a random sampling of the student's performance. (See "Bias in testing.")

Test item: The basic unit of content of an objective test.

Examples:

True-false item: $3 + 2 = 7$. True or False?

Multiple-choice item: Shakespeare wrote

- (a) The Iliad
- (b) King Lear
- (c) Tom Sawyer

Completion item: $3 + 2 = \underline{\hspace{2cm}}$

Matching items: Match each tool with the purpose for which it is ordinarily used:

<i>Purpose</i>	<i>Tool</i>
Cultivating ()	A. Needle
Sewing ()	B. Yardstick
Measuring ()	C. Hoe

Test item difficulty: The difficulty of a test item is defined in either of two ways, depending on the nature and purpose of the test.

1. If the test is norm-referenced, the difficulty of an item is usually defined in terms of the percent of students in the standard reference group who get the right answer. Thus, if 80 percent of the group answer the item correctly, the item would be regarded as "easy"; if only 20 percent got the right answer, the item would be regarded as "difficult."

2. In criterion-referenced tests, the difficulty of an item is usually defined in terms of the complexity of the thought process that the test maker thinks will be needed to arrive at the right answer. Thus $3 + 2 = ?$ would be judged an easy item; $3.764 + 2.005 = ?$ would be considered a harder item.

Test norm: An estimate of the average of the scores that would be found if a test had been taken under identical conditions by all the students in a clearly defined population of students. The estimate is usually obtained by testing under standard conditions a sample of the total population of students for whom the test is designed.

True score: The average of all the scores a student would obtain if he or she could be tested simultaneously with a very large number of tests, all made up from the same universe of tasks in a given area of subject matter. Thus, in the nature of the case, the true score is a theoretical concept. Even so, however, test makers have found various ways of getting, for any well-constructed test, a reasonably good estimate of the probability that any student's obtained score might be higher or lower than his/her true score by any given amount. (See "Standard error of measurement.")

Validity: The degree to which a test is relevant to some specific purpose which the test user has in mind.

APPENDIX C

Selected Readings

American Psychological Association and others. *Standards for Educational and Psychological Tests*. Washington, D.C.: American Psychological Association., 1974. 76 pp.

Latest enumeration of the standards that should govern the production, distribution, and use of educational and psychological tests, prepared by representatives from the APA, AERA, and NCME. Further recommendations in progress.

Anastasi, A. *Psychological Testing* (Fourth edition). New York: Macmillan Publishing Co., 1976. 750 pp.

Widely used textbook by an eminent psychologist, which deals clearly, authoritatively, and at length with the technical and practical aspects of testing in the schools and elsewhere.

Buros, O. K., ed. *The Eighth Mental Measurements Yearbook*. 2 Vols. Highland Park, N.J.: The Gryphon Press, 1978. 2182 pp.

Last in a long series of yearbooks containing critical reviews by experts of practically all educational and psychological tests available in the United States. Should be available in the professional library of every school district that makes any considerable use of tests in the schools.

Clasby, M. and others. *Laws, Tests, and Schooling*. Syracuse, New York: Syracuse University Research Corp., 1973. 209 pp.

A series of provocative papers dealing with federal and state actions with respect to local education. Included are such matters as statewide testing, testing for accountability, and testing for minimum competence.

NCCE/81

Cohen, M. D., ed. *Testing and Evaluation: New Views*. Washington, D. C.: Association for Childhood Education International, 1975. (unpaged)

Emphasizes the "qualitative" as distinguished from the "quantitative" approach to education evaluation as viewed by distinguished proponents of the open classroom.

Deighton, L. C., ed. *The Encyclopedia of Education*. Vols. 8 and 9. New York: The Macmillan Co. and the Free Press, 1971.

Contains a series of short articles on testing by various authors, each of whom is a recognized authority on the particular subject assigned to him or her. Titles include "Standardized Tests," vol. 8, pp. 391-394; and in vol. 9: "Test Information, Sources of," pp. 138-143; "Test Reliability," pp. 143-153; "Test Scores and Norms," pp. 153-165; "Test Validity," pp. 165-175; "Testing and Assessment Programs, National," pp. 175-179; "Testing and the Protection of Privacy," pp. 179-182; "Testing Programs, Statewide," pp. 182-185; "Testing, Projective Techniques," pp. 185-188; "Testing Situation," pp. 188-192; "Testing the Culturally Disadvantaged," pp. 196-201.

Dyer, H. S. *How to Achieve Accountability in the Public Schools*. Bloomington, Indiana: the Phi Delta Kappa Educational Foundation, 1973. 45 pp.

Philosophical discussion of the meanings of educational accountability and what is involved in making the concept work for the benefit of students and their schools.

Dyer, H. S. *The Interview as a Measuring Device in Education*. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, 1976. 16 pp.

"TM Report 56" from the Educational Resources Information Center. (A file of all such reports should be available in most school systems.) This one deals with the problems and possibilities of the face-to-face interview as a means of obtaining usable information about how students and their schools are faring.

Loewen, J. W., ed. *Testing and Citizenship*. Washington, D.C.: Clearinghouse for Civil Rights Research, Catholic University School of Law, 1978. 40 pp.

Double issue of the journal published by the clearinghouse, which in this case deals with recent developments in IQ research and the status of aptitude testing for admission to universities. Somewhat biased, but stimulating.

Turnbull, W. W. *Issues in Assessment of Basic Skills*. Princeton, N.J.: Educational Testing Service, 1978 (mimeo). 9 pp.

Witty discussion by the president of ETS concerning the "seven deadly sins" that threaten the usefulness of minimum competency testing.

Appendix D

Public Interest Principles for the Design and Use of Admissions Testing Programs

Proposed for discussion by:

*The College Board, Educational Testing Service,
Graduate Management Admission Council, Gradu-
ate Record Examination Board and Law School Ad-
mission Council.*

January 2, 1980

A number of the principles enumerated below have been cornerstones of most testing programs for some years. We believe it is important, however, to reaffirm them here to provide a fuller view of our beliefs and our expectations for the future.

Principles

1. We recognize the legitimate interest of the public in knowing what the tests contain and their efficacy in performing their intended functions. Therefore, we will implement the principle of publication of test content to a degree limited only by reasonable safeguards of efficiency, cost, quality, and the educational impact of the programs.
2. We fully support the principle of equity and we will continue to maintain and strengthen credible procedures for detecting bias and eliminating it from the con-

tent of the tests, while making such procedures visible to the public.

3. We recognize the need for routine procedures that allow the test taker to arrange for verification of the accuracy of the procedures determining the score attributed to him or her.
4. We believe that tests should be readily available to all individuals, regardless of conditions such as physical handicap or religious beliefs that may prevent the taking of exams under circumstances that meet the convenience of the majority.
5. We recognize that tests, together with the procedures for scoring them and reporting the results, should be designed to provide test takers with as much useful information as may be feasible about the specifics of their performance on the tests.
6. We reaffirm the right of individuals and institutions to privacy with regard to information by and about them, which should be safeguarded from unauthorized disclosure.
7. We recognize the need to formulate, maintain, and publish widely principles of appropriate use of scores and other test information derived from testing programs and to alert to and actively discourage misuse.
8. We recognize that both the institutions making use of test scores and the test takers themselves should have mechanisms through which to express their legitimate interests concerning the design and operation of testing programs and the use of the information derived from them.

Operational Elements

The separately constituted and governed groups sponsoring testing programs may choose to implement these principles in different ways. This probable diversity stems from differences in the nature and purposes of the tests in the several programs and from the specifics of their

structure and operation. Examples of possible approaches include the following:

1. Each prospective examinee should be able to receive a full-length sample of each test, similar to the one he or she will take, with the intended answers and with instructions for self-administration and self-scoring.
2. For tests given to a sufficient number of students annually to support the cost, at least one operational form of the test should be published periodically, in addition to the regular sample. A specific schedule of publication should be designated for each program.
3. Non-technical information about the testing program should be furnished routinely to test takers, users, and the general public. It should include a description of what each test measures, the error of measurement, how the scores are intended to be used, and a summary of the validity of the scores for the intended uses.
4. A technical publication should provide information on the same topics in sufficient depth to permit professionals in the field to assess the evidence and the accuracy of the non-technical summary.
5. Studies of the use of the test by professionals other than those in the sponsoring or administering agency should be actively encouraged and facilitated by provision of the necessary data with safeguards for individual privacy. The results of those studies should be published in regular journals and also incorporated in the technical and non-technical publications.
6. The test sponsor should ensure that operational forms of the test are independently reviewed before they are given. The review should include the appropriateness of the content of the test and in particular should seek to detect and remove potential racial, cultural or sex bias or other influences extrinsic to the characteristics, skills or knowledge to be measured. The review should also determine that the operational form is fairly represented by the sample test already distributed.

7. Test takers should have the right to question the accuracy of scoring, administrative procedures, specific questions in a test, or allegations of irregularities in test administrations. Current procedures to deal with this right should be reviewed and modified if necessary to ensure a fair and prompt response.

**PUBLICATIONS OF NATIONAL
COMMITTEE FOR CITIZENS
IN EDUCATION**

(All prices Include Postage and Handling)

H—PARENTS CAN UNDERSTAND TESTING

A nationally recognized testing expert helps parents understand the scores your child receives on tests at school. You'll find out how tests are used, how accurate scoring is, the controversies surrounding testing and what questions about tests you might profitably raise with your local school personnel. You will also learn what test makers' jargon means. (88 pages, \$2.50)

**I—THE SCHOOL BUDGET: IT'S YOUR MONEY;
IT'S YOUR BUSINESS**

This comprehensive guide tells parents and citizens where to get school budget documents, how to ask the right questions of school officials and how to act on the information you get to change budget priorities. (192 pages, \$4.95)

J—COLLECTIVE BARGAINING AND TEACHER STRIKES

The Expanding Role of Parents and Citizens. When teachers strike and the schools close, parents are uninformed spectators and school boards are too often unprepared for sophisticated contract negotiations. Should work contracts for public employees be modeled after private industry? Does the future hold an increased role in bargaining for parents and citizens? (88 pages, \$2.50)

K—WHO CONTROLS THE SCHOOLS?

Who has more power to decide how your child should be educated, the school board, the school administration, teacher organizations, the courts, or the federal government? Do parents and students have any power? Should they? Hard hitting, far-reaching conclusions and recommendations. (76 pages, \$2.50)

L—THE RIGHTS OF PARENTS IN THE EDUCATION OF THEIR CHILDREN

The most popular book ever published by NCCE tells parents in plain terms what rights exist under law for them and their children. The book's goal is to help parents resolve educational conflicts without going to lawyers or resorting to lawsuits. (162 pages, \$3.95)

M—DEVELOPING LEADERSHIP FOR PARENT/ CITIZEN GROUPS

What is leadership? How do you develop it in yourself and in others? How do you take the initiative? When do you assert yourself and when do you compromise? These and many other important questions about leadership development are addressed, using school-oriented problems as examples. (60 pages, \$2.50)

N—FUND RAISING BY PARENT CITIZEN GROUPS

Takes you through the basics, from identifying sources of support to the development of a fund-raising plan. Tells how to handle follow-up and cultivate donors. Includes sample proposals, suggestions on how to form a tax exemption, tips on reporting back to donors. (52 pages, \$2.50)

O—PARENTS ORGANIZING TO IMPROVE SCHOOLS

Step-by-step guide to organizing and running parent groups in your children's school that can act effectively to upgrade the quality of education and get parents into the educational scene in a lasting way. (52 pages, \$2.50)

P—LOS PADRES SE ORGANIZAN PARA MEJORAR LAS ESCUELAS

Spanish edition of Parents Organizing to Improve Schools. (64 pages, \$2.50)

Q—VIOLENCE IN OUR SCHOOLS: WHAT TO KNOW ABOUT IT—WHAT TO DO ABOUT IT

Includes regional surveys of school crime, do's and don't's for children's safety, how to conduct a survey to determine the level of security your community will support, plus much more. (52 pages, \$2.50)

R—NETWORK

The Paper for Parents—Eight issues during the school year keep you up to date on local and national happenings in education. Lets you exchange information and points of view with parents in other places. (Subscription, \$8 a year)

S—PARENTS' NETWORK DIRECTORY

Special Limited Edition—Describes goals and activities of over 200 parent/citizen groups coast to coast (\$2.50)

T—PARENT RIGHTS CARD

Wallet-size card lists 21 rights parents have in the education of their children under federal or state laws, regulations, and court decisions. Single card free if request is accompanied by self-addressed, stamped envelope. More than one copy, 10 cents per card. Discount on orders of over 250 cards. Also available in Spanish.

U—ANNUAL EDUCATION CHECKUP CARD

Guidelines to help parents review a child's progress in school. Fits in pocket or purse. Includes an

inventory of home and school files kept on school children, questions to ask teachers and guidance counselors, steps for appealing school decisions. Single cards free if accompanied by self-addressed, stamped envelope. Multiple copies at 25 cents per card. Discount of 10% for 100 copies or more. Also available in Spanish.

FILMSTRIPS

NCCE has produced a series of 5 slide-tape presentations based on Citizen Training Institutes and each running 12 to 15 minutes.

Individual filmstrips may be purchased for \$27.50, the full set of five for \$110.

Description of the 5 slide-tape presentations follows:

V—HOW TO WORK WITH SCHOOL OFFICIALS

Illustrates how to develop strategy for action, gather facts and make the parents' case to school officials.

W—FUND RAISING

Illustrates how to seek and obtain money from individuals, foundations, and other sources of funds.

X—ORGANIZING PARENTS AND DEVELOPING LEADERSHIP

Illustrates how to build a power base among parents and other citizens, run an organization, and develop effective leaders.

Y—PARENT INVOLVEMENT IN COLLECTIVE BARGAINING

Illustrates how collective bargaining process works and how parents can make their voice heard before, during, and after negotiations.

Z—THE LAW AND PARENTS' RIGHTS

Illustrates how current and proposed laws and court decisions affect parents' and students' rights.

Discounts Available on 10 or More Copies
USE ORDER FORM TO SPEED DELIVER
DETACH AND MAIL WITH CHECK OR
MONEY ORDER

PACKAGE PLANS

Used together, NCCE handbooks and related sound filmstrips can provide the basis for a successful parent/citizen training session. Filmstrips are based on actual training sessions for parents and are intended as introductions in each of five subject areas identified by parents as important. Handbooks provide in-depth background material. Each package is specially priced at \$29.95. If you order all five package plans (\$119.75) we will send you 25 copies of the Annual Education Checkup card free.

Package 1

Filmstrip/Tape:

Organizing Parents and Developing Leadership

Handbooks:

Parents Organizing to Improve Schools;

Developing Leadership for Parent/Citizen Groups

Package 2

Filmstrip/Tape:

How to Work with School Officials

Handbooks:

Parents Organizing to Improve Schools;

Developing Leadership for Parent/Citizen Groups

Package 3

Filmstrip/Tape:

Fund Raising

Handbook:

Fund Raising by Parent/Citizen Groups

NCCE/91

Package 4

Filmstrip/Tape:

The Law and Parents' Rights

Handbook:

The Rights of Parents in the Education of Their Children

Package 5

Filmstrip/Tape:

Parent Involvement in Collective Bargaining

Handbook:

Collective Bargaining and Teacher Strikes

HANDBOOKS IN THE PLANNING STAGES

SCHOOL BASED MANAGEMENT—NEW FAD OR NEW FORCE?

One of the most promising new practices for increasing involvement of parents and citizens in decision making at the local school building level is a decentralization process called school based or school site management. This handbook explains what it is, where it has been tried, what works and doesn't work, and how it can help citizens and school professionals share control of major school decisions.

THE IEP EXPERIENCE—FOR PARENTS OF HANDICAPPED CHILDREN.

In all parts of the country the parents of handicapped children are now required by federal law to be participants in the development of Individual Educational Programs. But NCCE's recent survey of thousands of these parents has indicated that the quality of parent participation needs to be greatly improved in many schools. A handbook for parents will tell how to get the most out of these joint planning sessions.

SINGLE PARENTS—WORKING WITH SCHOOLS

Single-parent families have unique needs as well as needs in common with two-parent families. This handbook, an outgrowth of a national survey of single parents, will tell how single parents and schools can work together. Methods for gathering information from local parents will be suggested. Programs and policies that are responsive to single-parent families will also be described.

**NATIONAL COMMITTEE FOR
CITIZENS IN EDUCATION
Suite 410 Wilde Lake Village Green
Columbia, Maryland 21044**

TO SPEED DELIVERY USE THIS ORDER FORM

NAME _____

ADDRESS _____

CITY _____

STATE _____

☐ Enclosed is my payment of \$ _____ for
publications listed below.

Publications:	QUANTITY	PRICE	TOTAL
The School Budget			
Parents CAN Understand Testing			
Who Controls the Schools			
The Rights of Parents			
Developing Leadership			
Fund Raising			
Parents Organizing to Improve Schools			
Collective Bargaining			
Los Padres Se Organizar			
Violence in Our Schools			
Parent Rights Card			
Annual Education Checkup Card			
NETWORK			

Filmstrips:	QUANTITY	PRICE	TOTAL
How to Work with School Officials			
Fund Raising			
Organizing Parents			
Parent involvement in Collective Bargaining			
The Law and Parents' Rights			

Package Plans:	Quantity	Price	Total
Package Plan 1			
Package Plan 2			
Package Plan 3			
Package Plan 4			
Package Plan 5			

DISCOUNT Orders of 10-19 copies deduct 10%, 20-49 copies, 15%,
and 50 or more copies 20%.

TOTAL: (Please Enclose Payment)

**NATIONAL COMMITTEE FOR
CITIZENS IN EDUCATION**

Governing Board

Katharine L. Auchincloss
NEW JERSEY
Chairperson

M. Hayes Mizell
SOUTH CAROLINA
Vice Chairperson

Peter Bittenwieser
PENNSYLVANIA
Secretary-Treasurer

Margaret P. Bates
CALIFORNIA

Katharine M. Bullitt
WASHINGTON

Lisle C. Carter
DISTRICT OF COLUMBIA

William K. Coblentz
CALIFORNIA

Frederick C. Green
DISTRICT OF COLUMBIA

Ernesto Loperena
NEW YORK

William J. McGill
NEW YORK

George Oser
TEXAS

Terry Sanford
NORTH CAROLINA

Nelson Schwab, Jr.
OHIO

Barbara O. Taylor
ILLINOIS

Mildred K. Wurf
DISTRICT OF COLUMBIA

SENIOR ASSOCIATES

Carl L. Marburger
J. William Rioux
Stanley Salett

Mitchell Rogovin
Rogovin Stern and Hugu
Counsel